Automatic detection and characterization of pulmonary nodules in thoracic CT scans

Colin Jacobs

This book was typeset by the author using LaTeX2$_\varepsilon$.

ISBN: 978-94-6259-880-5

Printed by Ipskamp Drukkers, Nijmegen.

Automatic detection and characterization of pulmonary nodules in thoracic CT scans

Proefschrift

ter verkrijging van de graad van doctor
aan de Radboud Universiteit Nijmegen
op gezag van de rector magnificus prof. dr. Th.L.M. Engelen,
volgens besluit van het college van decanen
in het openbaar te verdedigen op donderdag 19 november 2015
om 14.30 uur precies

door

Colin Jacobs

geboren op 24 augustus 1986
te Boxtel

Promotoren:	**Prof. dr. B. van Ginneken**
	Prof. dr. C.M. Schaefer-Prokop
Copromotor:	**Dr. E.M. van Rikxoort**
Manuscriptcommissie:	**Prof. dr. P.N.R. Dekhuijzen**
	Prof. dr. M.S. Brown (University of California, Los Angeles)
	Prof. dr. M. Oudkerk (Rijksuniversiteit Groningen)

The research described in this thesis was carried out at the Diagnostic Image Analysis Group, Radboud University Medical Center (Nijmegen, the Netherlands).
This work was funded by a research grant from MeVis Medical Solutions AG (Bremen, Germany).

Financial support for publication of this thesis was kindly provided by the Faculty of Science, Radboud University Nijmegen.

TABLE OF CONTENTS

1 Introduction **1**
1.1 Lung cancer . . . 2
1.2 Computed Tomography . . . 3
1.3 Lung cancer screening with CT . . . 5
1.4 Pulmonary nodules . . . 7
1.5 Computer-aided detection . . . 9
1.6 Lung nodule CAD . . . 10
1.7 Evaluation metrics . . . 11
1.8 Thesis outline . . . 13

2 Detection of subsolid nodules **15**
2.1 Introduction . . . 17
2.2 Materials . . . 18
2.3 Methods . . . 20
2.4 Results . . . 27
2.5 Observer study . . . 31
2.6 Discussion . . . 34
2.7 Conclusions . . . 38

3 Micronodule detection and quantification **41**
3.1 Introduction . . . 43
3.2 Methods . . . 44
3.3 Results . . . 47
3.4 Discussion and Conclusion . . . 49

4 Comparing CAD systems on a public database **51**
4.1 Introduction . . . 53
4.2 Materials and Methods . . . 54
4.3 Results . . . 58
4.4 Discussion . . . 62

5 Solid, Part-Solid, or Non-solid? **67**
5.1 Introduction . . . 69
5.2 Materials and Methods . . . 70
5.3 Results . . . 75
5.4 Discussion . . . 76

6 Detection of interval change **81**
6.1 Introduction 83
6.2 Methods 84
6.3 Results 90
6.4 Discussion 92

7 Integration into a screening workstation **97**
7.1 Translating research to the clinic 99
7.2 CIRRUS Lung Screening 101
7.3 Different versions of the software 106
7.4 Preliminary evaluation 106
7.5 Outlook 107
7.6 Conclusion 107

Summary **121**

General discussion **125**

Samenvatting **129**

Publications **133**

Bibliography **141**

Dankwoord **151**

Curriculum Vitae **155**

Introduction

1

This thesis focuses on the detection and characterization of pulmonary nodules on thoracic computed tomography scans. This first chapter provides a background for the other chapters in this thesis. First, we introduce some basic concepts of lung cancer, the technique of computed tomography (CT), lung cancer CT screening, and lastly the definition and type of pulmonary nodules will be described. After that, a general introduction into computer-aided detection is provided, followed by a more detailed section on computer-aided detection of pulmonary nodules. Then, the evaluation metrics used in this thesis are explained. Finally, an outline of this thesis is provided.

1.1 Lung cancer

Lung cancer is the most deadly cancer in both men and women worldwide[1]. It is the second most common cancer in both men and women, trailing prostate cancer for men, and breast cancer for women[1]. In the Netherlands, more than 10,000 people die of lung cancer every year[2]. The Dutch Cancer Society estimates that lung cancer will account for 25% of all cancer-related deaths in 2015[2]. In the United States, the American Cancer Society estimates a similar percentage for 2015: 27%[1]. Figure 1.1 shows the estimated number of new cancer cases and deaths in men and women in the United States in 2015.

The 5-year relative survival rate for all stages combined is only 17%[1]. This low rate can be largely attributed to the fact that at present, only 15% of all lung cancers are diagnosed in an early stage[1]. The reason for this is that symptoms usually do not occur until the cancer is in an advanced stage. If lung cancer is detected in an early stage when the disease is still localized and more curative treatment options are available, the 5-year relative survival rate is 54%[1]. Therefore, early detection of lung cancer is of major importance to reduce lung cancer mortality.

By far the most important risk factor for lung cancer is tobacco use. The risk increases both with quantity and duration of smoking. An estimated 87% of all lung cancer deaths are caused by cigarette smoking[3]. Therefore, complete banning of tobacco use would be the best recipe to reduce lung cancer mortality. Although the risks of smoking are well-known, it remains a major cause of the increasing global burden of cancer. Other risk factors for lung cancer are exposure to asbestos, exposure to radon, and air pollution.

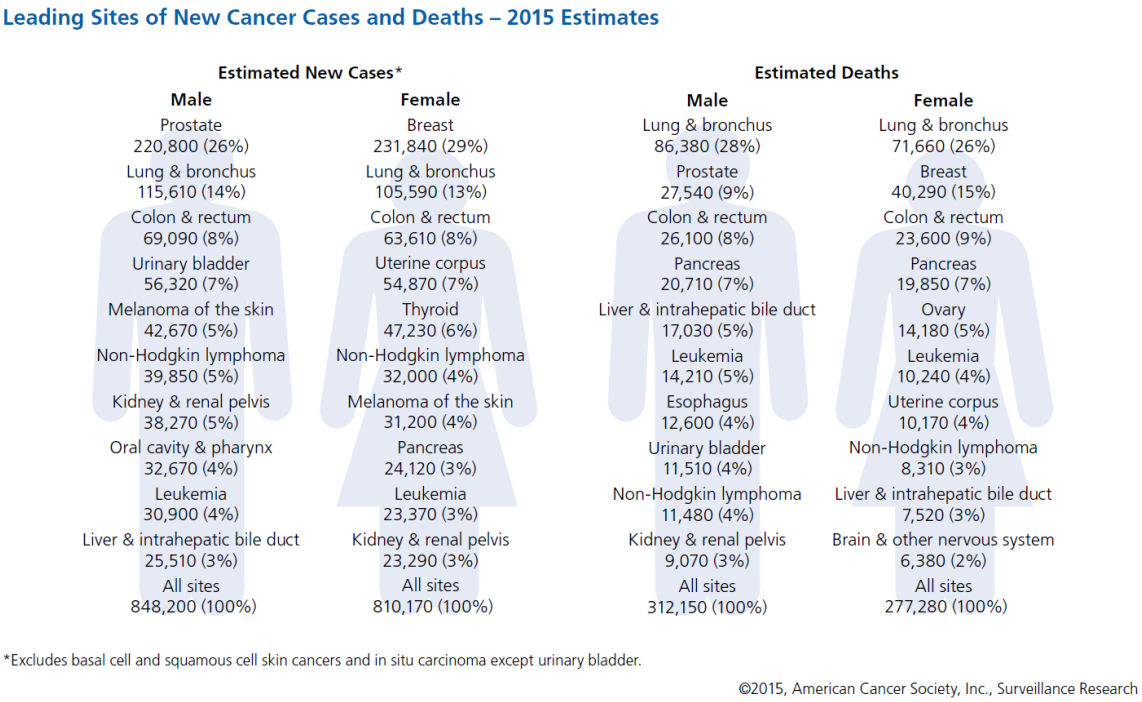


Figure 1.1: Estimates of new cancer cases and deaths for 2015. Image courtesy by the American Cancer Society.

1.2 Computed Tomography

Computed tomography (CT) is one of the most used imaging procedures in daily radiological practice. CT uses X-rays to acquire images. X-rays, also called Röntgen rays, are electromagnetic waves discovered by Wilhelm Conrad Röntgen in 1895[4]. Due to their penetrating ability, X-rays have been widely used in medicine to image the inside of a patient. In radiography, an object is placed in front of a X-ray detector. The object is illuminated with X-rays produced by an X-ray tube. The detector measures the amount of X-rays which have not been absorbed while passing through the object. Since certain structures absorb more X-rays than others, a projection image of the inside of the object is obtained.

During a CT acquisition, a thin axial section of a patient is imaged by taking a large series of two-dimensional X-ray projection images of this section from different directions. Using computer processing, a two-dimensional cross-sectional image of the scanned object, also called a slice, can be reconstructed from the projection images. In this way, many continuous axial slices can be obtained, which can then be stacked to form a three-dimensional image of the object. Each slice can be viewed individually to examine the whole inside of the scanned object.

In a CT scanner, an X-ray tube is rotated around the patient, with X-ray sensors

placed on the opposite side of the circle. In conventional CT scanners, the movement of the X-ray tube was limited and therefore, the X-ray tube had to be rewinded after each slice. So, after one slice was acquired, the tube had to be rewinded, the patient had to be moved along its axis and the next slice of the patient was acquired. In the 1980s, spiral CT scanning was introduced. In spiral CT scanning, the X-ray tube makes a continuous 360-degree movement while the patient is slowly moved through the scanner. Thus, the X-ray tube follows a spiral trajectory relative to the object without interruptions, making the acquisition of the CT scan faster. The next important development in CT scanners was multi-detector (or multi-slice) CT (MDCT) in which instead of a single row of detectors, multiple rows of detectors were placed such that multiple slices of the patient could be acquired simultaneously. State-of-the-art CT scanners have up to 320 rows of detectors, and are able to make a full thorax CT scan with a 1-mm axial spacing in a matter of seconds. This is well within the time most people are able to hold their breath.

The value of a pixel in a CT image represents the mean attenuation of the underlying tissue. Attenuation is measured on the Hounsfield scale, named after Sir Godfrey Newborn Hounsfield, one of the pioneers of CT imaging[5]. The Hounsfield scale defines the radiodensity of water at 0 Hounsfield units (HU), and the radiodensity of air at -1000 HU.

During a CT examination, the patient is exposed to radiation. People are exposed to radiation from natural sources such as naturally occurring radioactive materials and cosmic radiation from outer space all the time, and this is referred to as background radiation exposure. If the body absorbs large amount of radiation, there is an increased risk that cancer will develop. Small amounts of radiation are assumed to be harmful as well, although there are researchers which hypothesize that small amounts of radiation may be beneficial[6]. In medicine, the increased risk of developing cancer is considered to be greatly outweighed by the benefits of the examination. The amount of radiation which is absorbed by the body during the CT acquisition is referred to as the dose of the examination and is measured in millisieverts (mSv). CT scans can be made at different dose levels, which is determined by the voltage and current used in the X-ray tube. More dose typically means less noise on the CT scan. A typical clinical chest CT scan has an effective dose of around 7 mSv, and a low-dose chest CT scan is estimated to have an effective dose of around 1.5 mSv[7]. In comparison, a chest X-ray is estimated to have an effective dose of 0.1 mSv[7]. Recently, with the advent of iterative reconstructions, several studies have investigated the potential of using ultra low-dose CT, where the dose is comparable to a chest X-ray examination. For example, Nagatani et al. investigated nodule detection in chest CT scans obtained with mean effective doses of 0.29 mSv[8].

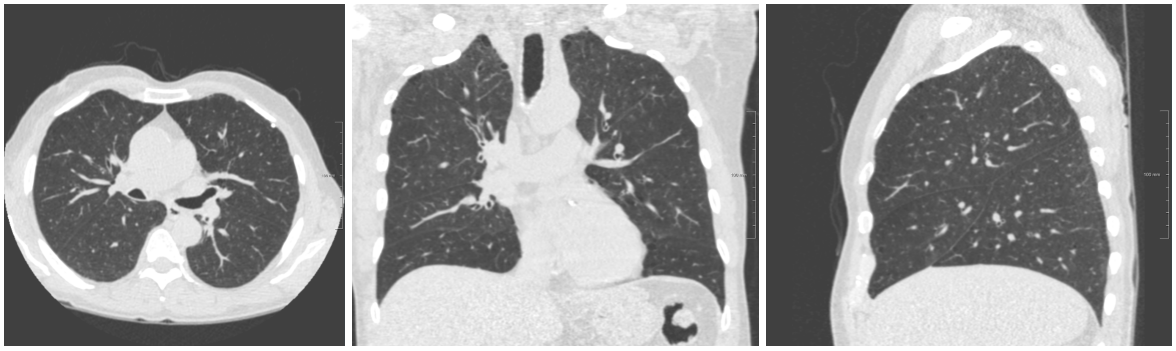


Figure 1.2: Three orthogonal images of a chest CT scan. Left: axial slice, middle: coronal slice, right: sagittal slice.

CT imaging of the thorax is superior to examine pathologies in the lung. Figure 1.2 shows a transverse, coronal and a sagittal slice of a thorax CT scan.

1.3 Lung cancer screening with CT

In the last two decades, lung cancer screening trials have been initiated to investigate whether mass screening of high-risk subjects would be beneficial in the fight against lung cancer. Since there exists a major risk factor for developing lung cancer, a group of high-risk subjects can be clearly defined: heavy smokers. By screening high-risk subjects, lung cancer can potentially be detected in earlier stage, thereby potentially reducing lung cancer mortality. Several randomized controlled trials (RCT) for lung cancer screening with low-dose computed tomography (LDCT) have been initiated. Low-dose chest CT produces images at sufficient quality to detect abnormalities in the lungs, but using significantly less radiation compared to a conventional chest CT scan. Table 1.1 gives an overview of the randomized controlled trials which have been conducted in Europe and the United States. The Dutch-Belgian NELSON lung cancer screening trial is the largest study in Europe, including 15,822 high-risk subjects[9]. NELSON was designed to investigate whether screening for lung cancer by low-dose multidetector CT in high-risk subjects will lead to a decrease in 10-year lung cancer mortality of at least 25% compared with a control group without screening. The National Lung Screening Trial (NLST) in the United States was the largest randomized controlled trial in the world, including 53,454 high-risk subjects[10]. The primary aim of NLST was to determine whether screening with low-dose CT, as compared with chest radiography, would reduce mortality from lung cancer among high-risk persons. Note that NLST used annual chest radiography in the control arm, and NELSON uses no screening in the control arm.

In November 2010, the initial findings of NLST were released; the NLST re-

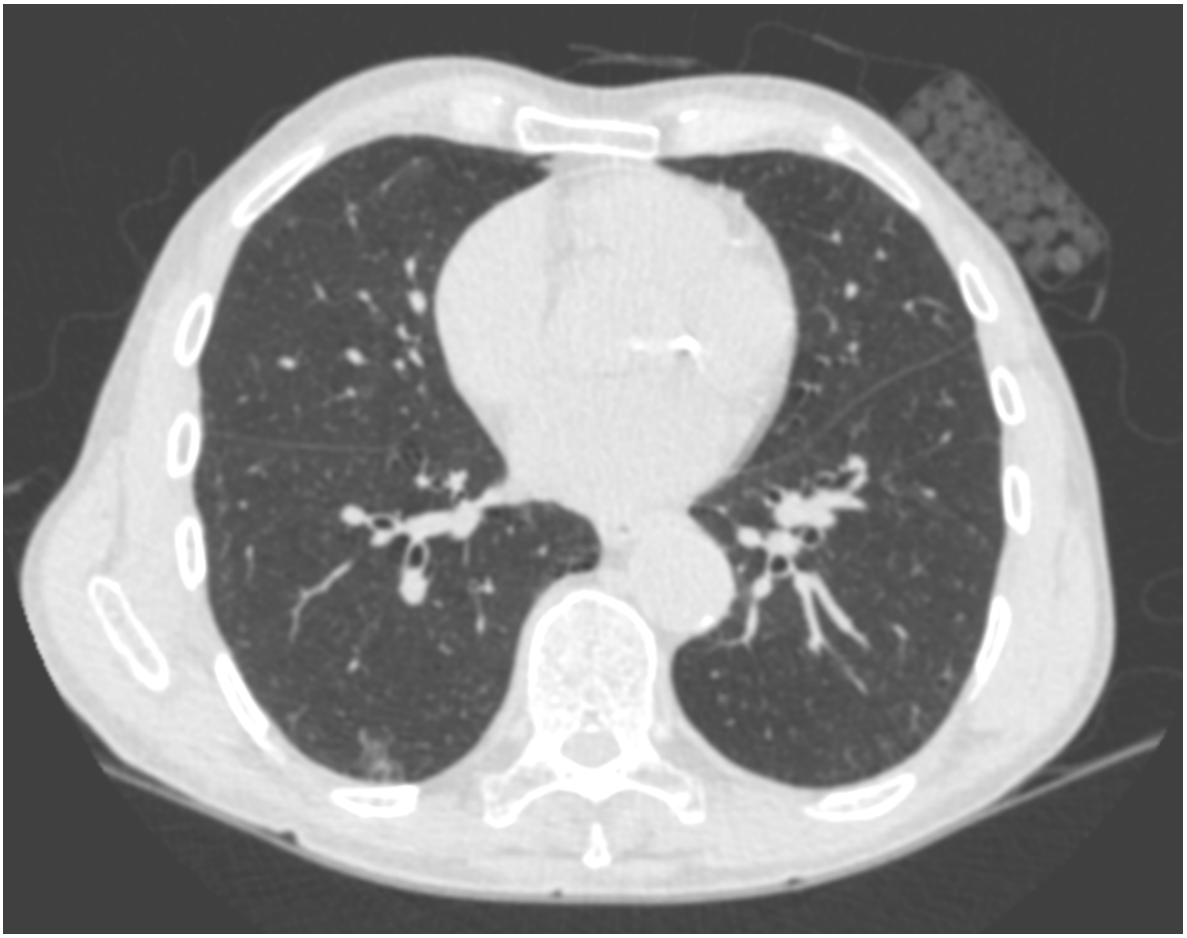


Figure 1.3: An axial slice of a thin-slice low-dose CT acquisition from the NELSON lung cancer screening trial. The pack of cigarettes which is visible in this examination illustrates the main risk factor for lung cancer: tobacco use.

searchers found a 20% lung cancer mortality reduction in their study group which received 3 annual rounds of low-dose CT screening in comparison to the control group, which received 3 annual rounds of chest x-ray screening[10]. This was the first RCT which showed clear scientific evidence that screening for lung cancer reduces lung cancer mortality. Based on this result, several organizations in the U.S. have started to endorse lung cancer screening using low-dose CT. In 2013, the U.S. Preventive Services Task Force (USPSTF) has given low-dose CT screening a grade B recommendation for high-risk individuals and early 2015, the U.S. Centers for Medicare and Medicaid Services (CMS) has approved CT lung cancer screening for Medicare recipients. This means that lung cancer screening is fully reimbursed by private insurance companies and Medicare in the U.S. for eligible subjects. A person is eligible to enter a lung cancer screening program when fulfilling the necessary eligibility criteria such as being asymptomatic, having an age between 55 and 77 years old, having a tobacco smoking history of at least 30 pack-years, and either being a cur-

Trial	Countries	# Participants	Groups
NLST	United States	53,454	LDCT, CXR
NELSON	Netherlands, Belgium	15,822	LDCT, controls
DLCST	Denmark	4,104	LDCT, controls
MILD	Italy	4,099	LDCT, controls
DANTE	Italy	2,472	LDCT, controls
ITALUNG	Italy	3,206	LDCT, controls
LUSI	Germany	4,000	LDCT, controls
UKLS	United Kingdom	4,000*	LDCT, controls

Table 1.1: Overview of the randomized controlled trials for lung cancer screening in Europe and the United States. Abbreviations: LDCT: low-dose CT, CXR: chest x-ray, NLST: National Lung Screening Trial, NELSON: Dutch acronym of the Dutch-Belgian Randomized Lung Cancer Screening Trial, DLCST: Danish Lung Cancer Screening Trial, MILD: Multicentric Italian Lung Detection, DANTE: Detection and Screening of Early Lung Cancer by Novel Imaging Technology and Molecular Essays, ITALUNG: Italian Lung, LUSI: German Lung Cancer Screening Intervention Trial, UKLS: U.K. Lung Screen. *The investigators from the UKLS trial are still recruiting for the pilot phase including 4,000 subjects. The main study will include 32,000 subjects.

rent smoker or quitted within the last 15 years.

The results of the NELSON trial, the second largest trial in the world, are pending; the first outcome data are expected to be published in 2016. Prompted by the situation in the United States and given the fact that no lung cancer screening recommendations existed in Europe, the European Society of Radiology (ESR) and the European Respiratory Society (ERS) have recently, based on the available evidence, provided new recommendations for lung cancer screening in Europe[11]. In their white paper, they recommend lung cancer screening in comprehensive, quality-assured, longitudinal programmes within a clinical trial or in routine clinical practice at certified multidisciplinary medical centres.

1.4 Pulmonary nodules

Interpretation of thoracic CT scans is a labor-intensive task for radiologists. Early stage lung cancer manifests itself as pulmonary nodules, which are described as round opacities, well or poorly defined, measuring up to 3 cm in diameter[12]. Detection of pulmonary nodules and accurate assessment of their risk of malignancy are crucial to find early stage lung cancer. Thin-slice helical chest CT scans have a sub-millimeter resolution at which small pulmonary nodules can be detected[13]. A

thin-slice CT acquisition of the thorax typically has a section thickness between 0.5 and 2 mm and hence generates between 200 and 500 transverse slices which all need to be inspected. Next to the large amount of slices, the rich structure of vessels and airways in the lungs complicates the search task even more. For this reason, there exists a substantial variability in the performance of human readers for detection of pulmonary nodules[14–16]. Next to the variability in the detection performance, there also exists a substantial interobserver variability among radiologists to what constitutes a pulmonary nodule[17–19], despite the seemingly clear definition[12].

Based on their appearance on CT, pulmonary nodules can be differentiated into subsolid and solid nodules[12]. Solid nodules have homogeneous soft-tissue attenuation on CT scans. Subsolid nodules can be further differentiated into non-solid nodules (synonym: ground glass nodules) and part-solid nodules (synonym: semi-solid nodules). Non-solid nodules manifest as focal areas of hazy increased attenuation that do not obliterate the bronchial or vascular margins. Areas of hazy increased attenuation are called ground glass opacity and therefore, these nodules are also referred to as ground glass nodules. Part-solid or semi-solid nodules contain both ground glass and solid components. In the Early Lung Cancer Action Project (ELCAP), 81% of all positive findings at baseline were solid nodules and 19% were subsolid nodules, which indicates that subsolid nodules are less common[20]. Although subsolid nodules are less common, this study published a significantly larger malignancy rate of 34% for subsolid nodules, compared to 11% for solid nodules[20]. Nodules are not only associated with lung cancer. In many lung diseases, nodular densities are seen in the lungs. An example is chronic silicosis, which is radiologically characterized by widespread, well-defined, small solid nodules, so called micronodules. These nodule have a diameter less than 3 mm. The high resolution of CT allows one to clearly differentiate micronodules. In this thesis, Chapter 3 is devoted to a dedicated detection system for micronodules. Although lung cancer may in a very early stage of their development be visible as a micronodule on CT, nodules below 3 mm are typically not annotated in a lung cancer screening setting because the a priori probability of such a nodule to represent lung cancer is well below 1%. Figure 1.4 shows an example of a solid, part-solid, and non-solid nodule.

The detection of pulmonary nodules receives most attention in this thesis, but also segmentation, characterization and tracking the evolution of nodules are essential to find early stage lung cancer. Accurate segmentation of pulmonary nodules is crucial to get reliable measurements of the volume, mass and average density of the nodule. Characterization is important to assess the likelihood of malignancy of nodules. Finally, growth is the most important property indicating the malignant character of a nodule and hence, tracking the evolution of nodules is of crucial im-

portance.

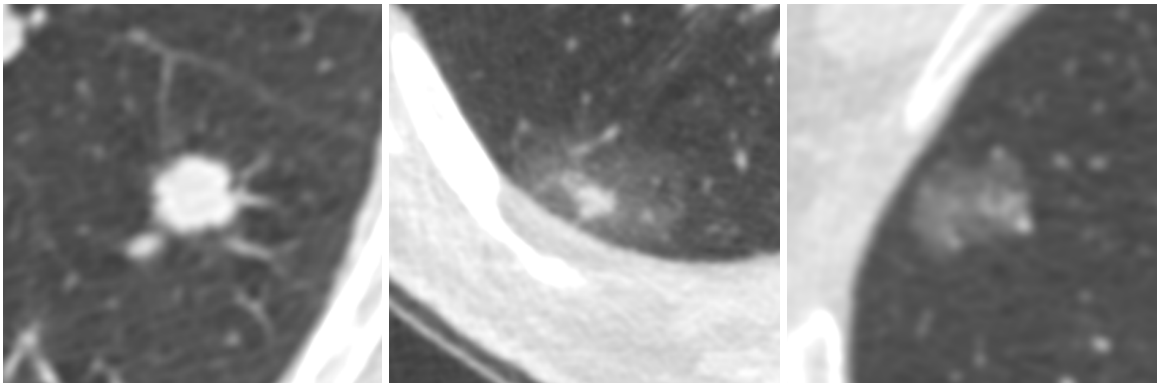


Figure 1.4: Different nodule types. Left: solid nodule, middle: part-solid nodule, right: non-solid nodule

1.5 Computer-aided detection

Computer-aided detection (CAD) is an extensive area of research in medical imaging and diagnostic radiology. CAD refers to technology which assists doctors in the interpretation of medical images. Due to the massive increase of imaging data produced by modern-day medical technology, radiologists have to analyze and evaluate more and more data in a relatively short time and CAD systems can potentially be of assistance. The acronym CAD has been used to represent both computer-aided detection and computer-aided diagnosis, which are also referred to as CADe and CADx. Computer-aided detection (CADe) algorithms are designed to assist doctors in *detecting* an abnormality, and thereby potentially decreasing observational oversights and thus the false negative rate of radiologists. Typical examples are detection of masses and microcalcifications in mammography, polyp detection in the colon, and the main focus of this thesis: lung nodule detection in CT. Computer-aided diagnosis (CADx) is technology which helps the radiologist in the *interpretation* of a radiologic finding, e.g. discriminating between benign and malignant breast lesions. The discrimination between CADe and CADx is not consistently used in literature; CAD sometimes refers to detection and sometimes to interpretation.

Integration of computer-aided detection systems into the reading process of radiologists can be accomplished in three different ways[21]:

1. First reader: After preselection by the CAD system, only the slices with CAD findings are presented to the radiologist.

2. Second reader: The radiologist reads the CT scan first without knowledge of the CAD findings. In a subsequent step he reviews the findings of CAD and decides if each CAD marking highlights a previously overlooked lesion or a false-positive finding.

3. Concurrent reader: The radiologist reads the CT scan with the CAD findings being displayed simultaneously. The radiologist can accept or reject the CAD findings and combine them with his/her own findings without the necessity of a second reading step.

The first CAD system was approved by the FDA in 1998 and indicated locations of potential breast cancer in mammograms. Commercial CAD systems which are approved by the Food and Drug Administration (FDA) are typically cleared for usage as a second reader. Thus, a radiologist first has to perform an initial read without the help of CAD.

1.6 Lung nodule CAD

Computer-aided detection of lung nodules has the potential to increase reader sensitivity for the detection of pulmonary nodules and may reduce reading time. The last two decades have shown substantial research into CAD for pulmonary nodules in thoracic CT scans[22,23]. Several studies have been conducted to investigate whether CAD can improve reader sensitivity for the detection of lung nodules[19,24–28]. In these studies, CAD is used as a second reader and showed improved detection rates. Although many academic and several commercial CAD algorithms have been developed, CAD for lung nodules is not commonly used in daily clinical practice. Possible explanations for this are a lack of reimbursement, technical impediments to integration into PACS systems, but also low sensitivity and high false positive rates. The positive results of the NLST lung cancer screening trial[10] and the subsequent developments towards implementation of lung cancer screening in the United States[29,30] have renewed the interest into CAD for pulmonary nodules. If lung cancer screening will be implemented on a large scale, the burden on radiologists will be substantial and CAD could play an important role in reducing reading time and thereby improving cost-effectiveness[31,32]. In the recently published white paper by the European Society of Radiology (ESR) and the European Respiratory Society (ERS), the authors explicitly recommend the use of computer-assisted nodule evaluation[11].

1.7 Evaluation metrics

This section presents the evaluation metrics used throughout this thesis.

ROC and FROC analysis

Receiver operating characteristic (ROC) analysis is a widely used methodology to measure the performance of a binary classification task. To explain the methodology, classification of images into diseased or not diseased is used as an example. When a radiologist classifies an image, he/she rates each image on a continuous scale, e.g. between 0 and 100, for having a disease. Then, by placing a threshold on the scores, a binary classification for benign and malignant can be obtained for each image. All images which are classified as having the disease are called the positives, and the images classified as not having the disease as negatives. The classification into positives and negatives is then compared with the reference standard and the images can be divided into the following four categories:

- True positive (TP): An image which is correctly classified as diseased.
- True negative (TN): An image which is correctly classified as not diseased.
- False positive (FP): An image which is classified as diseased, but is not diseased according to the reference standard.
- False negative (FN): An image which is classified as not diseased, but is diseased according to the reference standard.

Using these categories, *sensitivity* and *specificity* can be computed as formulated in Equation 1.1 and Equation 1.2. *Sensitivity*, also referred to as the *true positive rate*, measures the proportion of positives which are correctly classified as such. *Specificity*, also sometimes referred to as *true negative rate*, measures the proportion of negatives which are correctly classified as such.

$$\textit{Sensitivity} = \frac{TP}{TP + FN} \tag{1.1}$$

$$\textit{Specificity} = \frac{TN}{TN + FP} \tag{1.2}$$

An ROC curve can be created by plotting the sensitivity against the false positive rate (equal to 1-specificity). Different points of the ROC curve are obtained by applying different thresholds on the scores of the radiologist. For each radiologist, an

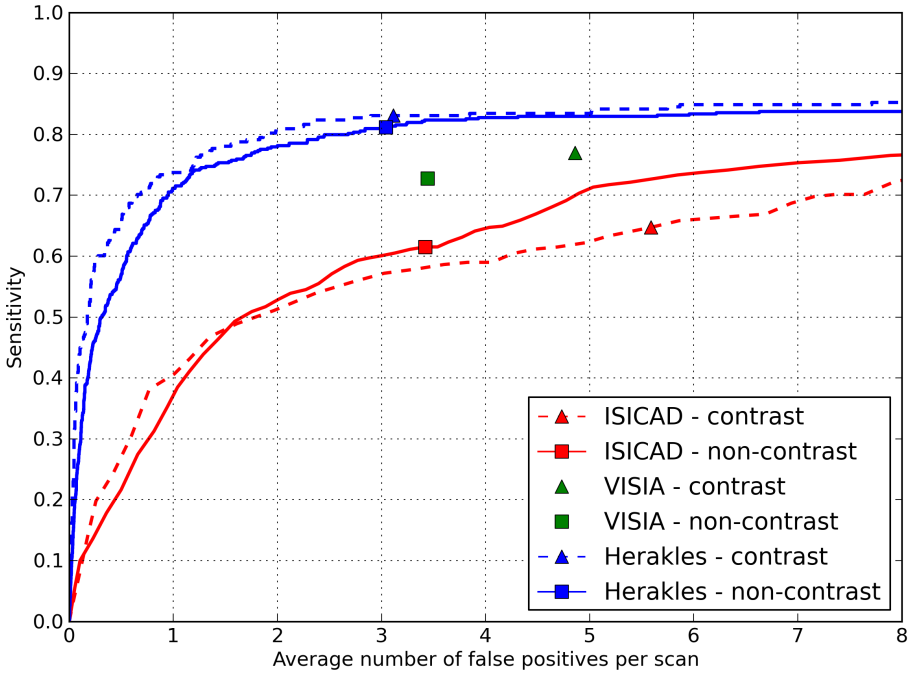


Figure 1.5: Examples of FROC curves, taken from Chapter 4 of this thesis. Different FROC curves for three CAD systems, ISICAD, VISIA, and Herakles, are shown for detection of nodules. For each system, two different datasets are use to evaluate: contrast scans and non-contrast scans. The points on the curves indicate the operating points of the three systems. For VISIA, no continuous FROC curve but only a single operating point can be provided since the CAD scores of this system are either 0 or 1.

ROC curve can be created and performance can be compared. Typically, the area under the ROC curve (AUC) is used as the figure of merit to compare performance. An AUC of 1 means perfect performance, and an AUC of 0.5 is equivalent to randomly guessing.

In this thesis, free-response operating characteristic (FROC) analysis is often used to analyze the performance of a CAD system. FROC analysis differs from ROC analysis in that there can be multiple locations of disease within an image and location information is associated with each abnormality. Thus, in the free-response paradigm, the radiologist or system does not know how many locations of disease there are or whether there are no locations of disease at all in an image. Whether a location in an image is detected by a system or radiologist is determined using a predefined hit criterium (for example, a CAD mark has to be located within 5 mm of an annotation made by a reference radiologist). In an FROC curve, sensitivity is plotted against the average number of false positive detections per scan. Fig. 1.5 shows an example with FROC curves, taken from Chapter 4 of this thesis.

Interobserver and intraobserver variability

In many radiological tasks, there exists a substantial variability in the performance of different readers. This is referred to as *interobserver variability*. Experienced radiologists may perform better than residents, for example. Next to that, one observer may have a different performance when he/she has to perform a certain task more than once on the same test set. This is referred to as *intraobserver variability*. In this thesis, Cohen κ statistics[33] were used to assess interobserver and intraobserver variability. Cohen κ statistics take into account the agreement that may occur by chance. Therefore, this is considered a better measure than simple percent agreement. Cohen's κ coefficient is calculated using Equation 1.3.

$$\kappa = \frac{p(o) - p(c)}{1 - p(c)}, \tag{1.3}$$

where $p(o)$ is the observed proportion of units in which the readers agreed, and $p(c)$ is proportion of units for which agreement is expected based on chance.

Thus, complete agreement will result in $\kappa = 1$ and complete disagreement will result in $\kappa = -1$.

1.8 Thesis outline

Automatic analysis of thoracic CT scans has the potential to improve detection rate of pulmonary nodules, reduce interobserver variability and speed up evaluation of screening CT scans. This however strongly depends on the performance of CAD systems. If the performance is good enough, it will probably positively influence the cost-effectiveness of lung cancer CT screening. This thesis presents novel algorithms to find abnormalities in thoracic CT images, which may be of benefit in the interpretation of chest CTs in lung cancer screening or clinical practice in general. The main focus of the thesis is automatic detection of pulmonary nodules in thoracic CT.

The outline of this thesis is as follows. **Chapter 2** describes a novel computer-aided detection system for subsolid pulmonary nodules. Detection of subsolid nodules has increased due to the use of thin-slice CT and the implementation of lung cancer screening trials and as a consequence, their prevalence and malignancy rate are better understood. Subsolid nodules are less common, but show a higher malignancy rate than solid nodules. We present a novel subsolid CAD system which is trained and validated with data from two sites from the NELSON lung cancer screening trial. In **Chapter 3**, a nodule detection system is developed for the detection of micronodules in subjects at high-risk for developing silicosis. Chronic silicosis is radiologically characterized by widespread, well-defined solid pulmonary

micronodules, measuring 3 mm or less. Early detection is crucial to stop progression but detection and quantification of these small nodules is tedious for human observers. We present an automatic method which finds micronodules and quantifies the micronodule load. **Chapter 4** describes a study in which we perform a comparative study of state-of-the-art nodule CAD systems on the largest publicly available reference data set, containing more than 1,000 CT scans. We perform an extensive analysis of the performance of three state-of-the-art CAD systems; two commercial and one academic CAD system. In **Chapter 5**, an automatic classification system for pulmonary nodules detected on CT is presented. Every nodule is classified as either solid, part-solid, or non-solid. This classification is crucial for management of pulmonary nodules. To put the performance of CAD into context, we also assess the interobserver variability among radiologists in this chapter. In **Chapter 6**, a system for automatic detection of interval change on consecutive CT scans is proposed. Interval change is of major importance when subjects are screened repeatedly. We propose an automatic detection system based on subtraction images which are acquired by employing an elastic registration method between consecutive CT images. **Chapter 7** discusses how the presented algorithms could be efficiently integrated into software that could be used in clinical routine or in CT lung screening programs.

Detection of subsolid nodules

2

C. Jacobs, E.M. van Rikxoort, T. Twellmann, E.T. Scholten, P.A. de Jong, J.M. Kuhnigk, M. Oudkerk, H.J. de Koning, M. Prokop, C. Schaefer-Prokop and B. van Ginneken

Original title: Automatic Detection of Subsolid Pulmonary Nodules in Thoracic Computed Tomography Images

Published in: Medical Image Analysis 18:374-384, 2014

Abstract

Subsolid pulmonary nodules occur less often than solid pulmonary nodules, but show a much higher malignancy rate. Therefore, accurate detection of this type of pulmonary nodules is crucial. In this work, a computer-aided detection (CAD) system for subsolid nodules in computed tomography images is presented and evaluated on a large data set from a multi-center lung cancer screening trial. The paper describes the different components of the CAD system and presents experiments to optimize the performance of the proposed CAD system. A rich set of 128 features is defined for subsolid nodule candidates. In addition to previously used intensity, shape and texture features, a novel set of context features is introduced. Experiments show that these features significantly improve the classification performance. Optimization and training of the CAD system is performed on a large training set from one site of a lung cancer screening trial. Performance analysis on an independent test from another site of the trial shows that the proposed system reaches a sensitivity of 80% at an average of only 1.0 false positive detections per scan. A retrospective analysis of the output of the CAD system by an experienced thoracic radiologist shows that the CAD system is able to find subsolid nodules which were not contained in the screening database.

2.1 Introduction

Subsolid pulmonary nodules occur less often than solid pulmonary nodules, but show a much higher malignancy rate. In the Early Lung Cancer Action Project (ELCAP), about half of the lung cancers found in this study originated from subsolid nodules[20]. Therefore, accurate detection of this type of pulmonary nodules is crucial. Although many systems have been proposed for detection of solid nodules (e.g.[34–44]), only few studies have focused on detection of subsolid nodules[45–49]. Kim et al.[49] described a slice-based CAD system using texture and intensity features. The system classified regions of interest (ROI) on manually chosen slices from CT examinations of 14 patients into ground glass opacity (GGO) or non-GGO. Performance was measured using receiver operating characteristic (ROC) analysis on all ROIs and showed an area under the ROC curve of 0.92. Zhou et al.[48] developed an automatic scheme for both detection and segmentation of subsolid nodules based on vessel suppression, intensity and texture analysis. They reported good performance but the test data set contained only 10 subsolid nodules. Ye et al.[47] presented a voxel-based method with rule-based filtering that was tested on 50 CT examinations with 52 subsolid nodules. They reported a high sensitivity of 92.3% but also a high false positive (FP) rate of 12.7 per scan. Tao et al.[46] developed a multi-level detection scheme with classification at voxel-level and object-level. They focused on small volumes of interest (VOIs) generated by a candidate detector algorithm which was not otherwise specified. The method was tested on a set of 1100 VOIs including 100 positive ones, from 153 healthy and 51 diseased patients. Results were provided for VOIs only, and neither the FP rate per scan nor the total number of VOIs per scan were reported. Finally, we published a preliminary version of our subsolid nodule CAD system which we trained and evaluated on a data set of 140 scans from one site of a large lung cancer screening trial. In this study, we reported promising results as we reached 73% sensitivity on the independent test set at an average of 1.0 FP/scan[45].

In this work, a novel automatic computer-aided detection system for subsolid nodules is presented and evaluated on a large database of a multi-center lung cancer screening trial.

The first algorithms for subsolid nodule detection have been evaluated on rather small amounts of data ranging from 10 to 50 thoracic CT scans containing between 10 and 52 subsolid nodules in total[47–49]. The most recent study used 1100 subvolumes from CT scans from around 200 subjects of which 100 subvolumes contained subsolid nodules[46]. In this study, we collected cases from two sites of a large lung cancer screening trial and included all CT examinations in which subsolid nodules

were annotated, leading to a larger database than presented by prior studies. Data collection is described in detail in Section 2.

The different components of the proposed CAD system, including candidate detection, feature calculation and classification, are described in Section 3. In comparison to the publication by Tao et al.[46], which is the most recent publication on subsolid nodule detection, we describe a full CAD system including the candidate detection step, which was not specified in the publication by Tao et al.[46]. Different classes of features have been utilized in previous publications. The first work by Kim et al.[49] and Zhou et al.[48] only used texture and intensity features. In the publications by Ye et al.[47] and Tao et al.[46], shape features are added to the feature set. In this work, we add another class of features by including context features which are calculated at the image-level, describing the relation of an area of ground glass opacity to its surroundings such as the lung, airways, vessels and other nodule candidates. We performed an experiment to show that these type of features benefit the classification performance. Experiments to optimize the configuration of the CAD system are also explained in Section 3. We experiment with many different classifiers to find the optimal classifier for this classification task. This is an important optimization of the CAD system which has also been mentioned as future work by prior publications[46].

Section 4 outlines the results of the CAD system on the independent test set. This section also reports the performance of a solid nodule CAD system on our database with many subsolid nodules and how the combination with a subsolid nodule CAD would benefit the detection performance for detection of subsolid nodules. At the best of our knowledge, we believe this has not yet been done in any publication to date.

In order to get a good insight into the performance of the CAD system and its potential to be applied in a clinical setting, an experienced thoracic radiologist reviewed the CAD marks of the CAD system on the independent test set and the results of this investigation are presented in Section 5. Finally, we discuss the performance and limitations of the proposed CAD system and opportunities for future work in Section 6 and conclude in Section 7.

2.2 Materials

Data for this study was collected from the NELSON trial, a large multi-center lung cancer screening trial, organized in the Netherlands and Belgium[9]. NELSON is an ongoing randomized control trial established to test if screening for lung cancer by low-dose CT in high-risk (ex-)smokers will lead to a 25% mortality reduction in lung cancer mortality.

In total, the trial includes 7557 participants who receive multiple rounds of screening with low-dose CT. All pulmonary nodules which were found during visual reviewing of the CT scans were recorded in the trial database. Among other characteristics, the screening radiologists indicated the location, diameter and nodule type of all detected pulmonary nodules; calcified, solid, part-solid or non-solid[9]. This database served as reference standard for our CAD analysis. We collected all thin-slice, low-dose CT examinations from two sites of the NELSON trial in which at least one subsolid (part-solid or non-solid) nodule was annotated. In total, the data set from which the scans were selected consisted of around 20,000 scans from around 4,500 subjects. All subsolid nodule annotations with a diameter smaller than 5 mm were discarded because current clinical guidelines state that these nodules do not require follow-up CT[50,51]. It has to be noted that in this work, solid nodule annotations were not included in the analysis.

The first site was the University Medical Center in Utrecht, The Netherlands. Screening CT examinations were made with a 16-detector row CT scanner (MX8000 IDT or Brilliance 16; Philips Medical Systems, Cleveland, Ohio) using a moderately soft reconstruction kernel (B; Philips Medical Systems). The second site was the Haarlemmer Kennemer Gasthuis in Haarlem, The Netherlands. At this site, CT examinations were made using a Somatom Sensation 16 (Siemens Medical Solutions) and reconstructed using a B30f kernel. All CT examinations at both sites were acquired in helical mode with 16 x 0.75 mm collimation. Exposure settings were 30 mAs at 120 kVp for patients weighing less than 80 kg and 30 mAs at 140 kVp for those weighing more than 80 kg. Axial images of 1.0-mm thickness were reconstructed at a 0.7-mm increment with a 512 x 512 matrix. In-plane voxel sizes varied from 0.53 mm to 0.89 mm.

After removal of the annotations of the subsolid nodules smaller than 5 mm, 209 scans from 103 patients were collected from the first site. In these patients, 122 subsolid nodules (63 part-solid, 59 non-solid) were found. Note that a pulmonary nodule can be annotated in multiple scans because follow-up examinations of a patient were included. Consequently, 225 annotations in 209 scans were found. This data set was solely used for training and optimization of the CAD system.

At the second site, 109 scans from 56 patients were collected after removal of the small subsolid nodules. In these patients, 60 subsolid nodules (32 part-solid, 28 non-solid) were found, which led to 114 annotations in the 109 scans. This data set, acquired with a different type of scanner from a different manufacturer, was used for independent evaluation of the CAD system.

The effective diameter of the nodules in both data sets varied between 5 and 34 mm, with a median of 10.7 mm.

2.3 Methods

This section describes the different steps of the CAD system. Prior to the detection pipeline, previously published lung, airway and vessel segmentation algorithms were applied[52–54]. An initial detection stage including nodule segmentation generates a set of candidates. A rich set of features was defined for subsolid nodule candidates. Previous publications on subsolid nodule CAD have used intensity, shape and texture features[46–49]. In this paper, we add another class of features by including context features. We performed classification experiments excluding and including context features to show the additional value of these features. We experimented with different classification schemes and different classifiers to investigate their influence on the classification performance. Previous publications already expressed the optimization of the classification as future work[46]. In this paper, we present an extensive and structured evaluation to select the best classification scheme. Based on the results of these experiments, the optimal configuration of the CAD system is chosen and this final system was evaluated on the independent test set. Finally, we combined our subsolid CAD system with a previously published solid nodule CAD system[36].

2.3.1 Candidate detection

Coarse candidate detection

The candidate detection procedure is started by applying a double-threshold density mask within the lung regions to obtain a mask of voxels with attenuation values commonly observed in ground glass opacities. A range between -750 and -300 Hounsfield units (HU) is used, similar to previous studies[55,56]. Partial volume effects at the edges of the lungs, vessels and airways can also give rise to attenuation values in the defined range. To remove these voxels, a morphological erosion operation using a spherical structuring element with a diameter of 3 voxels is applied. After this step, a connected component analysis is performed to cluster all voxels into candidates. Since subsolid nodules with a diameter smaller than 5 mm do not require follow-up CT[51], all candidates which have a volume smaller than 34 mm^3 (corresponding to the volume of an ideal sphere with a diameter of 4 mm are removed. Subsequently, a morphological dilation operation with the same structuring element is applied to undo the shrinking induced by the erosion operation. Finally, the volume and center of mass of all candidates are computed and candidates for which the centers of mass are within 5 mm of each other are merged. This merging procedure is applied to ensure that a nodule is covered by only one candidate.

Nodule segmentation

The candidate detection procedure described above generates clustered regions, but these are not an accurate segmentation of the subsolid nodules. Therefore, a previously published robust pulmonary nodule segmentation algorithm is used to acquire accurate nodule segmentations[57]. This algorithm works on a cubic volume of interest and applies an efficient combination of morphological operations to acquire a robust nodule segmentation. Automatic chest wall removal and separation from attached vasculature are incorporated into the algorithm. This method has been extensively evaluated for solid nodules and showed excellent results[57]. We slightly adjusted the segmentation algorithm to work for subsolid nodules. The algorithm described in the paper by[57] used a global lower threshold of -450 HU. In order to get good segmentation results for subsolid nodules (both part-solid and non-solid), the lower threshold is changed to -750 HU. The volume-of-interest for the segmentation algorithm is created around the center of mass of the candidate and the size of the VOI is set to 1.5 times the equivalent diameter of the initial candidate. In this way, accurate segmentations for all candidates are created and this forms the final set of candidates which go into the classification process.

2.3.2 Features

A rich set of features to describe the candidates is computed which can be subdivided into four categories: intensity, texture, shape features and context features.

Intensity features

Intensity features are calculated on four different sets of voxels:

- *segmentation*, voxels inside the candidate segmentation
- *boundingBox*, voxels inside a bounding box defined around the candidate segmentation
- *surrounding3*, voxels inside the surrounding of the candidate segmentation, created by dilating the candidate segmentation with a rectangular structuring element of size 3x3x3 voxels
- *surrounding5*, voxels inside the surrounding of the candidate segmentation, created by dilating the candidate segmentation with a rectangular structuring element of size 5x5x5 voxels

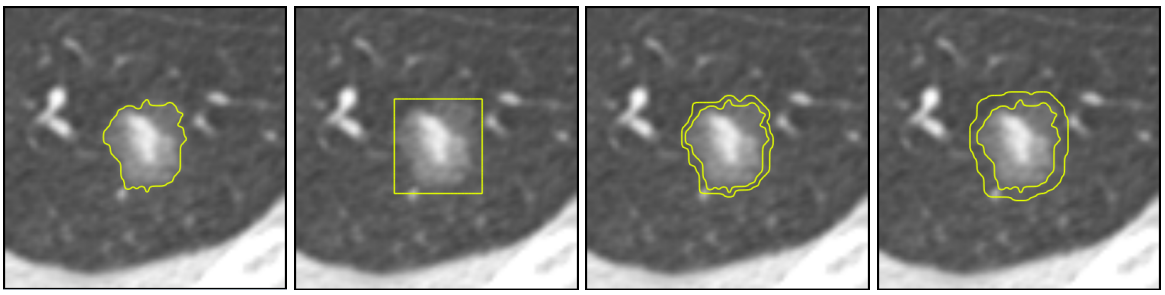


Figure 2.1: Four different sets of voxels on which features are calculated. Left: *segmentation*, Middle left: *boundingBox*, Middle right: *surrounding3*, Right: *surrounding5*.

An example of these regions is depicted in Fig. 2.1. These four regions are defined to extract features from the intensity profiles of the inner and the surrounding of a candidate. For each set of voxels, a normalized histogram is computed using a bin size of 50 HU. The bin size has been empirically determined such that the histograms are not too sparse, but still contain the necessary information to describe the underlying intensity distribution. For each normalized histogram, the following statistics are computed: entropy, mean, height of mean bin, mode (position of maximum peak), height of mode bin and value of the bins of the quantiles 5%, 25%, 50%, 75% and 95%. Furthermore, the standard deviation, minimum, maximum value and the first 7 Hu moments[58] are computed for voxel set *segmentation*. The Hu moments are translation, scale and rotation invariant and used to describe the underlying intensity profile. Finally, the maximum vesselness[59] over multiple scales (1.0, 1.77, 3.16, 5.62 and 10.0 voxels) is computed and the minimum, maximum, mean and standard deviation of the maximum vesselness in voxel set *segmentation* are used as features. Partial volume effects can create areas of ground glass opacity close to vessels or on vessel walls and by including these features, we capture information whether the candidate is in vicinity to or at a vessel wall.

In total, 54 intensity features are collected.

Texture features

For texture analysis, local binary patterns (LBP) and 2D Haar wavelets are used[60]. Both are commonly used texture descriptors to describe localized spatial texture information and have been used for parenchymal texture analysis in CT images[61]. These features for example help to exclude false positive candidates in regions of homogeneous ground glass opacity caused by motion artifacts. A VOI is created from the bounding box around the candidate segmentation and this volume is resampled (Lanczos resampling, $a = 3$) to two cubic volumes of interest of $16 \times 16 \times 16$ and $32 \times 32 \times 32$ voxels, respectively. Subsequently, 2D local binary patterns using a

neighborhood of 3×3 (P=8, R=1) are computed for every slice of the resampled volumes. Then, a normalized histogram with bin size 1 is computed of the LBP output for each volume and the same histogram statistics as mentioned in paragraph 2.3.2, except the quantiles, are used as texture descriptors.

Furthermore, 2D Haar wavelets are applied on the $32 \times 32 \times 32$ resampled volume. Each slice of the resampled volume is decomposed into four bands. Four volumes are constructed from the four bands from every slice. Three normalized histograms are computed from the three volumes built from the high-frequency bands. The volume built from the low-frequency band is not used. Again, the same histogram statistics are used as texture descriptors.

This leads to 40 texture features in total.

Shape features

The third group of features consists of shape features, which are computed from the candidate segmentation. Segmentation on other structures than nodules can create odd shapes and therefore, shape is an important feature to discriminate true positive from false positive samples. First, the following features are calculated: *sphericity*, *compactness1*, *compactness2* and *guessRadius*. In order to calculate the *sphericity*, a sphere S is defined at the center of mass of the candidate region with the same volume as the candidate segmentation. Then, *sphericity* is defined as the ratio between the volume of the voxels of the candidate segmentation within sphere S and the total volume of sphere S. Then, in order to calculate *compactness1*, *compactness2* and *guessRadius*, the bounding box around the candidate segmentation is used and the dimensions are named *dimx*, *dimy* and *dimz*. To calculate *compactness1*, the number of voxels of the candidate cluster is divided by the total number of voxels within the bounding box. *Compactness2* is calculated by dividing the number of voxels in the candidate cluster by the number of voxels in a cube for which the size is defined by the largest dimension of the bounding box ($max(dimx, dimy, dimz)$). The feature *guessRadius* is calculated by dividing the volume of the bounding box by 6. In case of a perfect spherical nodule, this will produce the exact radius of the sphere. Secondly, the number of voxels and the cluster size in mm^3 are computed to describe the size of the candidate. These two features are almost identical, but the cluster size takes the resolution of the CT scan into account. Finally, the same set of 7 invariant Hu moments are computed from the candidate mask voxels to describe its shape. Note that in contrast to the previous calculation of Hu moments for the intensity features, the voxels are in this case set to 1 inside the segmentation and 0 outside the segmentation.

In total, 13 shape features are computed.

Context features

Finally, a novel group of context features is defined, which describe the location of the candidate region in respect to the lung boundary, the airway tree, the vessels and other subsolid nodule candidates. The location of the candidate with respect to the lungs, vessels and airways is important for multiple reasons. For example, larger areas of ground glass opacity can be seen at the gravity dependent portions of the lung (base of the lung when CT is performed supine) due to microatelectasis. This will result in a candidate which has an elongated shape along the boundary of the lung. A combination of shape and context features is able to capture this. Another example is airways filled with mucus, which can manifest in the intensity range of ground glass opacities. These candidates will show an overlap with the airway segmentation and this can be used to classify them as false positive. Furthermore, the relation of candidates to other candidates is relevant contextual information. For example, a small candidate which is surrounded by many other candidates is more likely to be originating from an area of microatelectasis than to be a subsolid nodule.

First, two distance transforms are calculated within the lung regions; the first using the lung segmentation and the second using the airway tree. The distance to the lung boundary and distance to the closest airway is extracted from the distance transforms for all voxels inside the candidate segmentation. The mean, standard deviation, minimum and maximum distance to the lung boundary and airways are computed and used as context features.

Secondly, a bounding box is defined around the lungs and this is used to compute relative position features; relative X, Y and Z position, and distance to left bottom corner of the bounding box are computed. Furthermore, the distance to the center of mass of both lungs is calculated.

Thirdly, the absolute and relative airway and vessel overlap are computed. To calculate this, we count the number of voxels within voxel set *boundingBox* which are part of the airway segmentation or the vessel segmentation. The exact number of voxels is the absolute overlap and the relative overlap is calculated as the number of voxels inside the segmentation divided by the total number of voxels in *boundingBox*.

Finally, the relation of a candidate with respect to other candidates is described. First, we use the total amount of candidates within the scan as a feature. This provides information about the number of ground glass areas in the lung. Secondly, we calculate the number of candidates within a distance of 30 mm and 50 mm of the candidate and the distance to the closest other candidate.

In total, this sums up to 21 context features.

2.3.3 Classification

In this section, the experiments to optimize the classification performance are described. Furthermore, we describe the evaluation of the CAD system on the independent test set. During evaluation of the system, a nodule is marked as detected when the center of mass of the candidate is within a distance R of the center of the nodule. In order to ensure that the CAD mark is displayed within the nodule on the CT scan, we set R to be the radius of the nodule size. This radius is half of the diameter which is reported by the radiologist during reading of the CT in the screening.

Optimization of the classification scheme

In order to select the best classification scheme, several classification experiments are conducted. These experiments are performed in 10-fold cross-validation on the training set. Since patients can have multiple scans of the same pulmonary nodule, the folds are created by splitting at a patient level to prevent bias. Candidates are classified into two classes: nodule or false-positive (FP), and the final performance of the CAD system is evaluated using free-response operating characteristic (FROC) analysis.

Classification of candidates is tested using a single versus a two-stage classification scheme. The one-stage classification scheme computes the complete set of features for all candidates and uses one supervised classifier to classify all candidates into the two classes. In contrast, the two-stage classification scheme utilizes only five features in the first stage to perform a first-stage classification. This first-stage classification is aimed at removing as many false-positive candidates as possible. Then, the complete set of features is only calculated for all remaining candidates. This two-stage approach has two advantages. Firstly, the computation time of the CAD system for the two-stage classification scheme will be shorter since the complete set of features does not have to be calculated for all candidates. Secondly, the first-stage classification could make the data set more balanced, which could be beneficial for the classifiers tested in the second stage. Both approaches are tested to evaluate which classification scheme is optimal in terms of classification performance.

The first stage classification of the two-stage classification scheme is performed using a Linear Discriminant Classifier (LDC)[62] because of its simplicity and speed. The optimal set of five features for this first-stage classification is determined using three approaches. In the first two approaches, a sequential forward floating selection (SFFS) procedure[63] is used to select 5 features. The SFFS procedure uses a random 50% of the training fold as training data and the other 50% as testing data. In the first approach, accuracy is used as optimization criterium for the SFFS procedure.

In the second approach, the partial area under the FROC curve between 0 and 3 FP/scan is used as optimization criterium. Finally, in the third approach, the Fisher's linear discriminant ratio[64] is calculated for all features and the five features with the highest ratio are selected. Note that this approach does not consider feature combinations. Three different LDC classifiers are trained using these three different sets of five features and this produces a likelihood for all candidates for each classifier. The likelihood threshold for stage one for each classifier is determined by sorting all likelihoods of the positive samples and selecting the lowest likelihood. Consequently, no true positives are removed in the training set. All candidates with a likelihood below this threshold are removed. In this way, we obtain the set of five features which removes the most false positive candidates without removing true positives.

For the one-stage classification scheme and the second stage of the two-stage classification scheme, a k-nearest neighbor classifier (kNN)[65], random forest classifier (RF)[66], GentleBoost classifier (GB)[67], nearest mean classifier (NM)[62], support vector machine using radial basis function kernel (SVM-RBF)[68], and LDC are tested in order to find the optimal classifier for this classification task. Parameters of the different classifiers were also optimized in cross-validation on the training set. In the kNN classifier, K was set to the square root of the number of positive samples. The random forest classifier was trained with 100 trees with a maximum tree depth of 20. For the GentleBoost classifiers, regression stumps were used as weak classifiers and 250 weak classifiers were used to train the classifier. Next to these classifiers, a combination of 10 GentleBoost classifiers, referred to as GB10, is used. This classifier consists of 10 GentleBoost classifiers, which are all separately trained using a random 75% of the training set. The final output of the GB10 classifier is the median of the 10 different classifier probabilities. In pilot experiments, this classifier produced improved performance compared to a single GentleBoost classifier. For the GB10, we also used 250 regression stumps in each separate GB classifier. The C and gamma parameter of the SVM-RBF classifier were optimized in an inner 5-fold cross validation loop within the training fold of the 10-fold cross validation loop. As an optimization criterium, the partial area under the FROC curve between 0 and 3 FPs/scan was used. All features are normalized to zero mean and unit variance.

Evaluation of benefit of the context features

We hypothesize that the presented context features contribute to a significantly better classification performance. We conducted an experiment on the training set using the final system with and without the context features to test this hypothesis. The bootstrap method is used to test statistical significance. Scans were sampled with replacement from the cross validation set 5000 times. Every bootstrap sample had

the same number of scans as the original data set. Classification performance is measured by the partial area under the FROC curve between 0 and 8 FPs/scan.

Evaluation of the optimal classification scheme on independent test set

The optimal classification scheme, one-stage or two-stage classification, and the optimal classifier was chosen based on FROC analysis of the cross-validation results on the training set. Then, the optimal classification scheme was trained using the complete training set and tested on the independent test set to evaluate the performance of the CAD system. Note that this test set has not been used in any way during the optimization of the classification scheme.

2.3.4 Combination with solid nodule CAD

In clinical practice, the subsolid nodule CAD system will operate in combination with a solid nodule CAD system. Although solid nodule CAD algorithms are not optimized and trained for detection of subsolid nodules, they may still detect a fraction of all subsolid nodules. In particular, they may be sensitive to detecting the solid core of part-solid nodules. Therefore, the combination of a solid nodule CAD and the proposed subsolid nodule CAD may increase the overall detection sensitivity of subsolid nodules. To evaluate this, a previously published nodule CAD system[36] is applied to all cases in the test set. This CAD system reached an excellent sensitivity in a large comparative study of nodule CAD algorithms, the ANODE09 study[69]. In this work, the CAD system was set to operate at an average of 4 false positives per scan. Note that this operating point has been determined on an independent data set so the system is not guaranteed to generate precisely this false positive rate on our test set.

2.4 Results

2.4.1 Candidate detection

The candidate detection step generated 237 $\pm$ 267 candidate regions per scan in the training set and 109 $\pm$ 127 candidate regions per scan in the test set. In the training set, the candidate detection sensitivity was 84% for all subsolid nodules, where the sensitivity for part-solid nodules and non-solid nodules separately was 81% and 87%, respectively. In the test set, the sensitivity was 88% for all subsolid nodules, and 85% and 90% for part-solid and non-solid nodules, respectively.

2.4.2 Classification

Optimization of the classification scheme

The FROC curves of the different classifiers in the single stage classification scheme on the full training set in 10-fold cross validation are depicted in Fig. 2.2. This figure shows that the GB10 classifier performs best and reaches 69% sensitivity at 1 FP/scan and 74% sensitivity at 2 FPs/scan. Note that the candidate detection sensitivity is 84%, which means that the classification sensitivity cannot be higher than this value. This is indicated by the dotted line in Fig. 2.2.

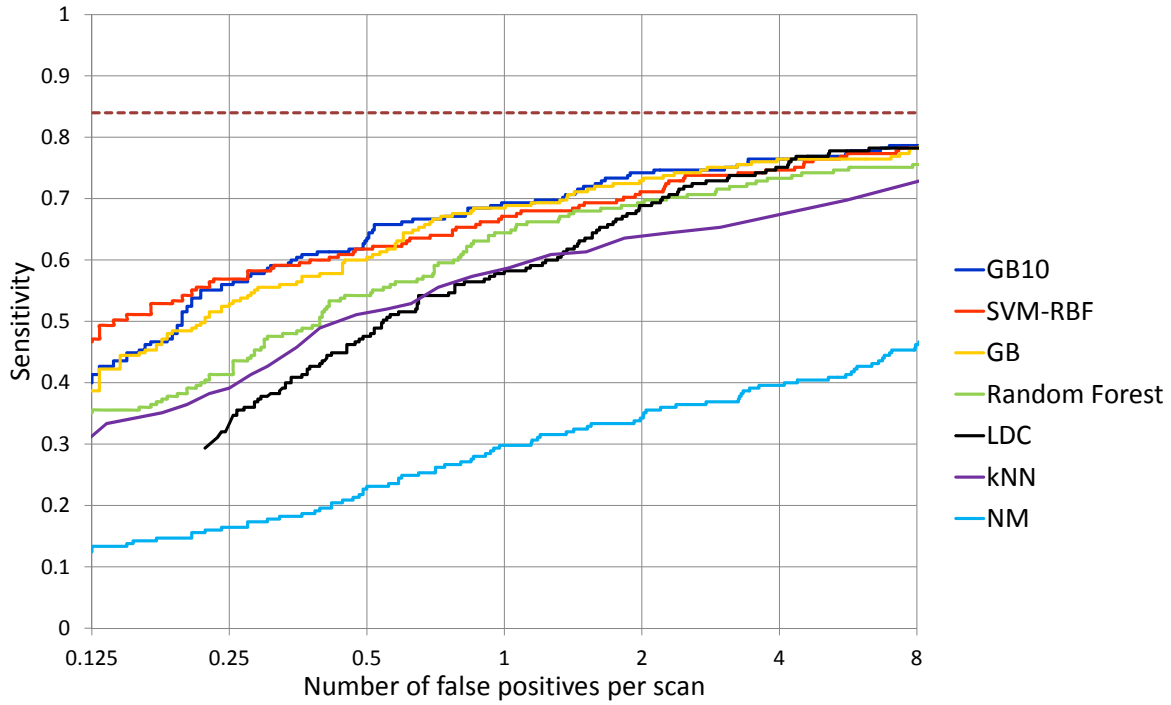


Figure 2.2: FROC curves of the different classifiers in a single stage classification scheme tested in 10-fold cross-validation on the full training set. The horizontal axis is logarithmic. The dotted line indicates the maximal sensitivity which can be reached due to the candidate detection sensitivity. GB10: combination of 10 GentleBoost classifiers, as explained in Section 2.3.3, GB: GentleBoost classifier, SVM-RBF: Support vector machine with radial basis function kernel, LDC: linear discriminant classifier, kNN: k-nearest neighbor classifier, NM: nearest mean classifier.

The goal of the first stage classification is to reduce the amount of FPs in the set of candidate regions. Using the three approaches explained in section 2.3.3, three different sets of five features for stage 1 classification are constructed and their performance is tested in 10-fold cross-validation. Table 2.1 shows the amount of samples which are removed when a posterior probability threshold T is used which removes

Feature set	Reduction ratio
SFFS - Accuracy	**68%**
SFFS - Partial Area under FROC curve	**79%**
Fisher Linear Discriminant Ratio	**59%**

Table 2.1: Performance of the first-stage LDC classifier for different feature groups selected by: (a) sequential feed-forward selection (SFFS) using accuracy as optimization criterium, (b) SFFS using partial area under the FROC curve between 0 and 3 FPs/scan as optimization criterium, (c) Fisher's linear discriminant ratio (FLDR). Right column shows the reduction ratio: the percentage of remaining candidates after removal of candidates below the likelihood threshold.

no true positives in the training set. This table shows that the feature set based on Fisher's linear discriminant ratio removes the most samples and therefore, this feature set is selected for the first stage of CAD system.

The performance of the different classifiers in the second stage is depicted in Fig. 2.3. Comparable to Fig. 2.2, the GB10 classifier also performs best in this second-stage classification. Again, note that the candidate detection sensitivity is 84%, which means that the classification sensitivity cannot be higher than this value.

Evaluation of the optimal classification scheme on independent test set

The performance of the best classifier for the single stage classification scheme, GB10, is compared to the best configuration for the two-stage classification scheme; LDC with the FLDR features and the GB10 for the second stage. There is no significant difference between the two classification schemes.

Computation time of the system with different classification schemes was measured on a Dell laptop using 1 core with a processor speed of 2.70 GHz. Using the one-stage classification scheme, the average computation time per case was 129 seconds. Using the two-stage classification scheme, the average computation time per case was 122 seconds. These results exclude the computation time of the lung, airway and vessel segmentation. Although the computation time difference is small, the two-stage classification scheme is selected as the final classification scheme. The complete configuration of the CAD system is depicted in Fig. 2.4.

Finally, Fig. 2.5 shows the FROC curve of the performance of the final configuration of the CAD system on the test set. This system has a two-stage classification scheme using a LDC and GB10 classifier, is trained on the complete training set and tested on the test set. Furthermore, FROC curves for part-solid nodules only and for non-solid nodules only are displayed. This figure shows that the final configuration

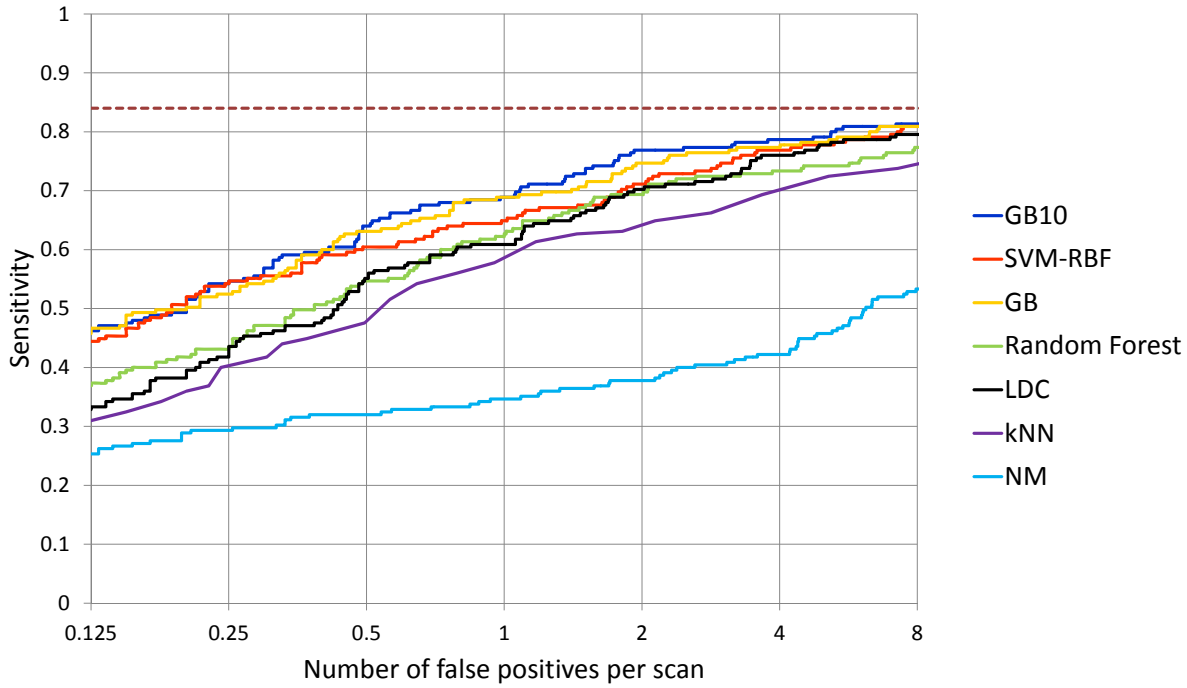


Figure 2.3: FROC curves of the different classifiers in a two-stage classification scheme tested in 10-fold cross-validation on the full training set after first stage classification. The horizontal axis is logarithmic. The dotted line indicates the maximal sensitivity which can be reached due to the candidate detection sensitivity. GB10: combination of 10 GentleBoost classifiers, as explained in section 2.3.3, GB: GentleBoost classifier, SVM-RBF: Support vector machine with radial basis function kernel, LDC: linear discriminant classifier, kNN: k-nearest neighbor classifier, NM: nearest mean classifier.

of the CAD system reaches 80% sensitivity at 1 FP/scan and 83% sensitivity at 2 FPs/scan. Note that the candidate detection sensitivity on the test set was 88% (85% for part-solid nodules and 90% for non-solid nodules). After the first stage classification, three true-positives (TP) were removed from the test set. These were all non-solid nodules. Consequently, the maximal classification sensitivity in this graph is 85% (97 out of 114) for all subsolid nodules, 85% (46 out of 54) for part-solid nodules and 85% (51 out of 60) for non-solid nodules. Note that the CAD system is able to detect all subsolid nodules when the system is set to operate at 4 FPs per scan.

Evaluation of benefit of the context features

Fig. 2.6 shows the FROC curves of the performance of the CAD system on the training set with and without context features. In addition, to get an impression of the value of each feature group, we also added the FROC curves of the system when only

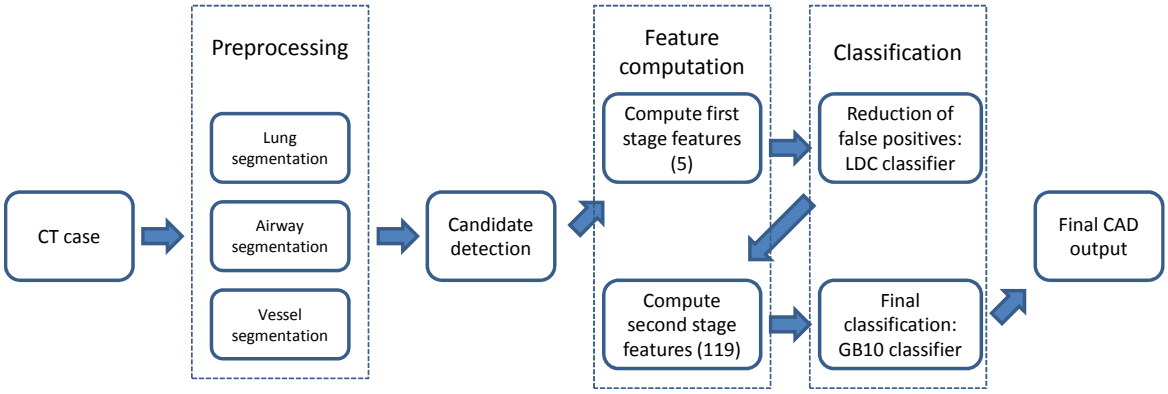


Figure 2.4: Schematic diagram of proposed CAD system.

one of the feature groups is used. The performance of the CAD system, measured by the partial area under the FROC curve between 0 and 8 FPs/case, is significantly increased when the context features are added to the system ($p = 0.001$). The FROC curves of the single feature groups also confirm the value of the context features as they are the most important feature group after the intensity features.

2.4.3 Combination with solid nodule CAD

The complete test set has been processed using the solid nodule CAD system, which was set to operate at an operating point of 4 false positives per scan. At this operating point, the solid nodule CAD detected of 55% of all subsolid nodules in the test set. The detection rate for part-solid nodules and non-solid nodules was 71% and 42%, respectively. Furthermore, the solid nodule CAD detected 9 (6 part-solid and 3 non-solid) of the 23 subsolid nodules which were missed by the subsolid CAD system when operating at 1.0 FP/scan. Consequently, adding the solid nodule CAD system to the subsolid nodule CAD system increased the detection sensitivity for subsolid nodules from 80% to 88%.

2.5 Observer study

In this section, we describe an analysis of the output of the CAD system using an independent experienced chest radiologist as reference. Prior publications have shown that databases used for evaluation of CAD have its limitations[69,70]. Although the process of obtaining a reliable reference standard is a widely known problem when evaluating CAD systems, it is important to investigate the output of a CAD system and evaluate whether additional lung lesions are detected[71]. The annotations of the

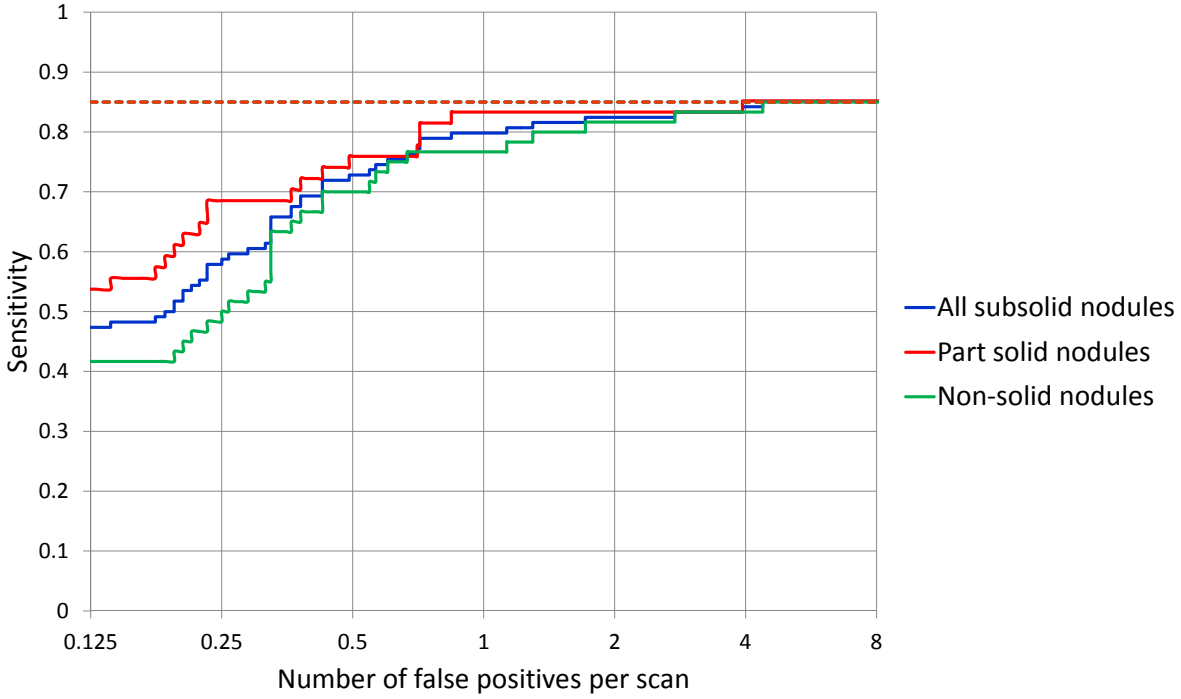


Figure 2.5: FROC curves of the CAD system on the independent set for all subsolid nodules, part-solid nodules only and non-solid nodules only. The horizontal axis is logarithmic. The dotted lines indicate the maximal sensitivity which can be reached due to the candidate detection sensitivity for the different nodule groups.

database used for this study were obtained without the support of CAD and therefore, it is likely that CAD identified lesions not annotated in the screening trial. It remains open whether these nodules have been overlooked by the radiologists during screening or whether they were deemed not suspicious enough to be annotated, for example in case of small subsolid nodules. In order to see if CAD find additional lesions and to get an idea of the nature of false positives which were generated by the CAD system, all false positives of the CAD system when set to operate at an average of 1 FP per scan are selected for evaluation by the radiologist.

Prior research has shown that there is a lot of debate among radiologists as to what constitutes a pulmonary nodule. This is clearly illustrated by the study of the Lung Image Database Consortium (LIDC)[70]. The database of this study is publicly available and contains 1018 thoracic CT scans. In this study, four experienced thoracic radiologists first independently reviewed the CT cases and recorded the nodules they found; the initial blinded-read phase. They categorized the nodules into three categories: "nodule≥3mm", "nodule<3mm", "non-nodule≥3mm". Then, in a second session, the radiologist could see the anonymized annotations of the other three radiologists and gave a final opinion about the nodules in the CT case. After

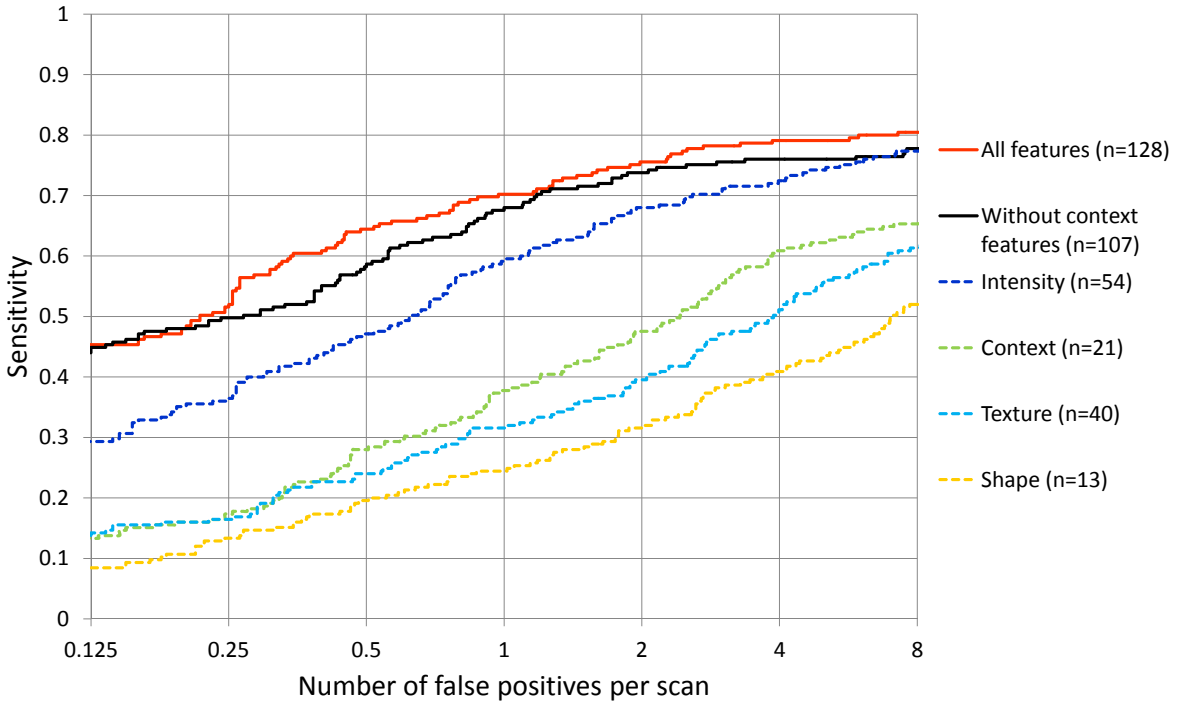


Figure 2.6: FROC curves of the CAD system with and without context features. In addition, the FROC curves when the CAD system is trained with only one single feature group are depicted. The FROC curves are obtained by a 10-fold cross-validation on the full training set. The horizontal axis is logarithmic.

both sessions, 2669 lesions were marked as "nodule≥3mm". Of these, only 34.8% were marked by all radiologists and 29.1% were marked by a single radiologist only. Since the investigators of LIDC did not force consensus, they were able to show that radiologists disagree considerably about what constitutes a pulmonary nodule. In order to test whether there are debatable subsolid nodules within the annotated nodules of the test set, the independent radiologist was also confronted blindly with all subsolid nodules from the reference standard.

In total, the radiologist was thus confronted with 223 marks in random order; 109 false positives and 114 nodule annotations from the reference standard. An in-house developed, dedicated reading workstation was used in which the radiologist could easily navigate from mark to mark and inspect findings in all orthogonal planes simultaneously. All reading facilities of a usual reading workstation were present. Note that the radiologist read these marks in a blinded fashion. Based on experience from prior experiments, the following categories were defined: subsolid nodule, subsolid nodule <5mm, solid nodule, scar, acute inflammation, interstitial lung disease (ILD), motion/pulse artifacts and other FP. The last category contains everything not belonging to the first groups and the radiologist could provide a comment with the

mark type	subsolid nodule	subsolid nodule <5mm	solid nodule	inflammation	scar	interstitial lung disease	motion/pulse artifacts	other FP	**total number**
True positive	75	0	2	0	6	0	0	8	91
False negative	10	0	4	0	2	0	0	7	23
False positive	44	6	1	0	5	2	4	47	109

Table 2.2: Analysis of the true positives, false negatives and false positives of the CAD system at 1FP/scan by an experienced thoracic radiologist. All marks are categorized into one of the 8 categories, represented by 8 columns in the table.

mark. In addition, when a subsolid nodule was marked, the radiologist was also asked to indicate whether it was a part-solid or non-solid nodule. The results from this observer study are presented in Table 2.2.

2.6 Discussion

A fully automatic CAD system for automatic detection of subsolid nodules was described and extensively evaluated. The data set used in this work was collected from two sites of a large multi-center lung cancer screening trial and was considerably larger than the data sets used by previous publications[46–48]. CAD systems trained and tested on data from the same scanner may show better performance than the actual performance of the CAD system on data from different scanners. To circumvent this problem, we decided to test the CAD system on data from a different site, which was acquired with a different scanner.

The CAD system is initiated with a lung, airway and vessel segmentation algorithm. These algorithms were successful on all scans from the training set and failed on only one scan of the test set. Consequently, no marks were generated on this scan in the test set and lesions on this scan were missed. Evaluation and improvement of these algorithms is beyond the scope of this paper.

The candidate detection procedure is started in the next step in which a robust and accurate nodule segmentation algorithm is integrated. As a basis for a robust performance of the CAD system, a sensitive candidate detection step is essential. As the results in Section 2.4.1 show, the sensitivity of the candidate detection step is

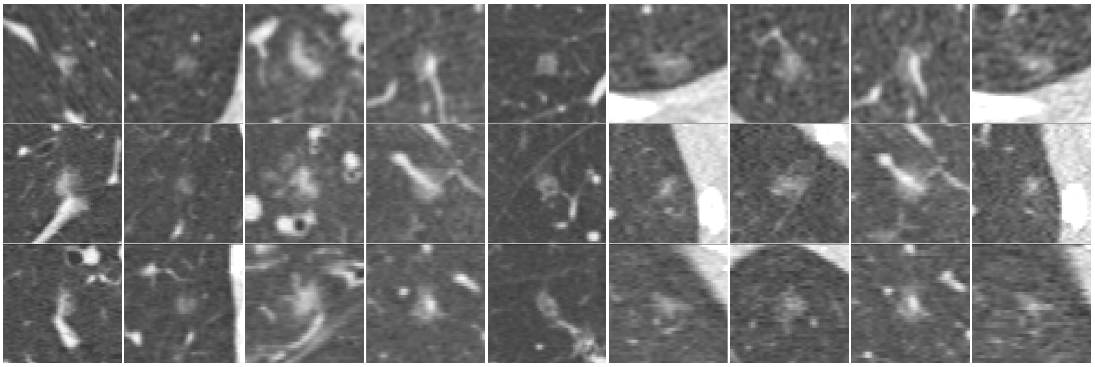


Figure 2.7: Subsolid nodules detected by CAD and confirmed by radiologist, but not present in reference standard. Every column shows one nodule. The top image displays an axial image, the middle image displays a sagittal image and the bottom image a coronal view. Images show a field of view of 40x40mm in which the nodule is centered.

high; 84% on the training set and 88% on the independent test set. We believe this is a sufficiently high sensitivity for the candidate detection step.

Subsequently, a rich set of features is defined for the candidate regions. In addition to intensity, shape and texture features, we added a novel group of context features. Contextual information is commonly used in classification problems in many fields. However, despite the large number of publications on CAD, contextual information is barely used in this area, with few exceptions such as the work of Sánchez et al.[72], Song et al.[73] and Hupse and Karssemeijer[74]. Hupse and Karssemeijer[74] and Song et al.[73] used specific contextual features for detection of masses in mammograms and detection of tumors and lymph nodes in thoracic images, respectively. Sánchez et al.[72] presented a general framework for including contextual information and showed that this significantly improved the classification of two CAD applications: identification of exudates and drusen in 2D retinal images and coronary calcifications in 3D computed tomography scans. The contributions of our paper in terms of contextual classification are twofold. Firstly, we introduced novel context features dedicated to subsolid nodules and their relation to vasculature, pleural surface and the bronchial tree. We show that these features significantly increase the classification performance ($p = 0.001$). Secondly, compared to Sánchez et al.[72] we introduced a new class of context features which take the context of the complete image into account instead of only the context of individual candidates. The rationale behind this new class of features is that in a scan with many subsolid nodule candidates, it is more likely that the high number of candidates is caused by other factors such as interstitial lung disease, low inspiration level or poor image quality than that there are actually many subsolid nodules present within the scan.

In order to get the best performance out of all features, two different classification schemes and several different classifiers have been tested. To prevent a positively biased performance on the test set, the complete optimization of the CAD system is performed in 10-fold cross-validation on the training data. The FROC curves in Fig. 2.2 and Fig. 2.3 show that classification of candidate regions using all features achieves the best performance when the GB10 classifier is used. Therefore, this classifier was used in the final configuration of the CAD system. There was no clear difference in performance between the two classification schemes. Since the two stage classification scheme reduces the computation time of the CAD system, this is used in the final system. Although the reduction in computation time is only minor at this point, the benefit of the two stage approach might increase when the feature set is extended.

The lack of performance difference between the one stage and two stage classification scheme using the GB classifier is an interesting observation which shows that the GB classifier intrinsically is able to handle unbalanced data sets. Boosting classifiers increase the weights of misclassified examples in each iteration step of the training and therefore, many obvious negative samples will soon get a low weight which effectively balances the data set. Therefore, no large improvement can be found for the GB classifier when comparing the one-stage versus two-stage classification. On the contrary, a substantial improvement in the two-stage classification situation can be seen for the LDC and NM classifier, which are known to suffer from unbalanced data sets. In the LDC for example, the estimate of the common covariance matrix of the two classes is a weighted mean of the two sample matrices. Therefore, it will be dominated by the variation of the prevalent class. Consequently, a substantial bias may be present if the assumption of a common covariance matrix does not hold. CAD researchers typically invest considerable amount of time in optimizing the classification scheme and one aspect of this is to start experimenting with one stage versus multiple stage classification schemes. Next to that, different classifiers are usually tested. We show that the possible performance increase of a two-stage classification is dependent on the classifier, since certain classifiers are able to handle unbalanced data better than others. This should be taken into account while developing CAD algorithms.

In clinical practice, the subsolid nodule CAD system will operate in combination with a solid nodule CAD system. Most published and commercial nodule CAD systems generate between 2 and 4 false positives per scan. In order to control the false-positive rate of the combined system, good sensitivity at a low false-positive rate is desired. The FROC analysis in Fig. 2.5 shows that the proposed subsolid CAD system reaches a sensitivity of 80% on our independent test set at a low false-positive

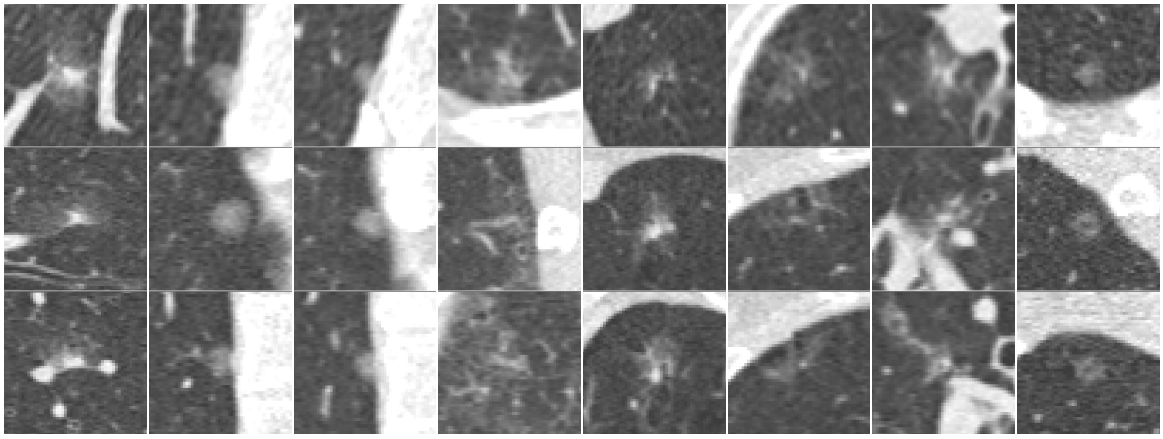
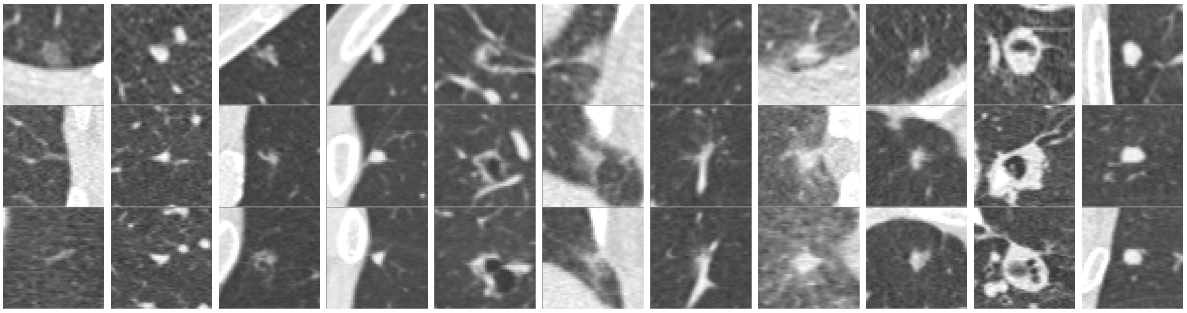


Figure 2.8: Subsolid nodules missed by subsolid nodule CAD when set to operate at 1 FP/scan. Every column shows one nodule. The top image displays an axial image, the middle image displays a sagittal image and the bottom image a coronal view. Images show a field of view of 40x40mm in which the nodule is centered. Also note that the second and third nodule is the same lesion but at different time points.

rate of 1 FP/scan. Given that the system will be combined with a solid nodule CAD, we believe that this is a suitable operating point of our subsolid nodule CAD system. As Fig. 2.5 also shows, the performance for part-solid nodules is better than for non-solid nodules. The most likely reason for this is that the solid component in the part-solid nodule is well described by our feature set and this benefits the classification process.

Furthermore, we explored whether a solid nodule CAD system would be able to detect the missed nodules of the subsolid CAD system. At an average of 1 FP/scan, the subsolid nodule CAD reaches a 80% sensitivity which means that it misses 23 of the 114 annotations in the test set. As described in Section 2.4.3, the solid nodule CAD detected 9 of the 23 missed subsolid nodules when set to operate at 4 FPs/scan. Consequently, the sensitivity would increase from 80% to 88% which represents an important improvement. To get an idea which type of nodules we still miss, 7 of the 18 missed subsolid nodules are depicted in Fig. 2.8.

Finally, an experienced thoracic radiologist scored all findings in the reference standard as well as all CAD marks when the CAD is set to operate at an average of 1 FP/scan. Table 2.2 shows that 44 from the 109 (40%) false positives CAD marks were retrospectively marked as subsolid nodules. These nodules were initially not annotated during the screening. These results show that a subsolid nodule CAD system may detect subsolid nodules overseen by radiologists. In Fig. 2.7, 9 of these false-positive CAD marks which were retrospectively classified as subsolid nodule are shown. Furthermore, Table 2.2 also shows that from the 23 missed subsolid nodules,



(a) (b) (c) (d) (e) (f) (g) (h) (i) (j) (k)

Figure 2.9: Subsolid nodules in the reference standard which are not classified as subsolid nodules by the independent radiologist. Every column shows one nodule. The top image displays an axial image, the middle image displays a sagittal image and the bottom image a coronal view. Images show a field of view of 40x40mm in which the nodule is centered. The categories which were chosen by the experienced radiologist for these nodules were: (a) other FP: plate atelectasis, (b) other FP: perifissural nodule (PFO), (c) other FP: small complex lesion next to bullae, (d) other FP: PFO, (e) other FP: small complex lesion next to bullae, (f) scar, (g) scar, (h-k) solid nodule

a considerable amount (13) were not marked as subsolid nodule by the radiologist. This indicates that these nodules are debatable. In Fig. 2.9, examples of these 13 debatable nodules are shown.

In future work, we intend to further explore how to optimally combine the proposed subsolid nodule CAD system with a solid nodule CAD and evaluate this on a large data set containing both solid and subsolid nodules. Then, robust detection of the complete spectrum of pulmonary nodules is at hand.

2.7 Conclusions

In this work, a fully automatic CAD system for detection of subsolid nodules was presented. The data set used in this work was collected from a large multi-center lung cancer screening trial and is considerably larger than the data sets used in previous studies. A novel set of context features is introduced which describe the relation of a nodule candidate to the lung boundary, airways, vessels and other nodule candidates. Using experiments, we have shown that these features significantly improve the classification performance of the CAD system. The CAD system reached 80% sensitivity on the independent set at an average of 1.0 false positive per scan. We believe this is an appropriate operating point because in clinical practice, a subsolid nodule CAD will be used in combination with a solid nodule CAD, which usually

generates between 2 and 4 FPs per scan. When the subsolid nodule CAD system was combined with a previously published solid nodule CAD, the sensitivity for subsolid nodules increased to 88% on the independent test set. An extensive evaluation of the CAD output using an experienced thoracic radiologist showed that a substantial fraction of the false positives of the system when operating at 1.0 false positive per scan were actually considered to be subsolid nodules which were missing in the reference standard.

Acknowledgments

We thank the investigators of the NELSON trial for providing data for this study. Funding for this research was provided by MeVis Medical Solutions AG.

Micronodule detection and quantification

3

C. Jacobs, S.H.T. Opdam, E.M. van Rikxoort, O.M. Mets, J. Rooyackers, P.A. de Jong, M. Prokop and B. van Ginneken

Original title: Automated detection and quantification of micronodules in thoracic CT scans to identify subjects at risk for silicosis

Published in: Medical Imaging, volume 9035 of Proceedings of the SPIE, page 90351I, 2014

Abstract

Silica dust-exposed individuals are at high risk of developing silicosis, a fatal and incurable lung disease. The presence of disseminated micronodules on thoracic CT is the radiological hallmark of silicosis but locating micronodules, to identify subjects at risk, is tedious for human observers. We present a computer-aided detection scheme to automatically find micronodules and quantify micronodule load. The system used lung segmentation, template matching, and a supervised classification scheme. The system achieved a promising sensitivity of 84% at an average of 8.4 false positive marks per scan. In an independent data set of 54 CT scans in which we defined four risk categories, the CAD system automatically classified 83% of subjects correctly, and obtained a weighted kappa of 0.76.

3.1 Introduction

Silicosis is an incurable lung disease and one of the oldest known occupational diseases: The ancient Greeks and Romans were already aware that breathing dust may cause respiratory illness. Chronic silicosis is radiologically characterized by widespread, well-defined solid pulmonary micronodules, measuring 3mm or less[12]. In an advanced state, this may be seen on chest radiographs, but micronodules are much better visible, and probably in a much earlier stage already detectable on computed tomography (CT) scans. It is difficult to diagnose silicosis unequivocally, and therefore a combination of a history of dust exposure, radiological manifestations and exclusion of other diseases is used[75]. Even then, silicosis is difficult to recognize, especially in early phases of the disease where symptoms are often absent, or resemble those of chronic obstructive pulmonary disease. Early detection is however, very important because there is no cure and ending exposure is the only way to avoid progression.

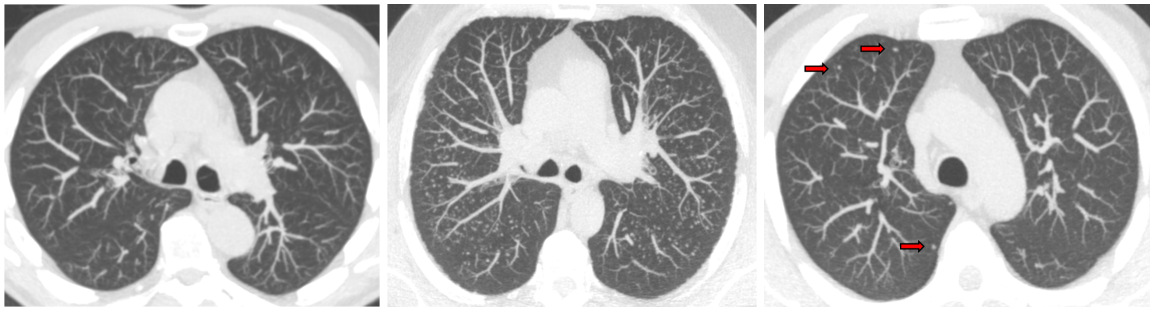


Figure 3.1: Three axial sections which illustrate the detection task for micronodules. All images are axial maximum intensity projections of 10 mm. The left scan is a normal subject without micronodules. The second scan shows disseminated micronodules, compatible with radiological features of manifest silicosis. In the right image, a subject which had 14 micronodules is depicted. The arrows indicate the three detected micronodules in this particular section of the scan. Evidently, the detection of a small amount (> 13) of micronodules to detect early stage silicosis is a tedious task.

In a recent study, Mets et al.[76] compared CT scans of two groups of 54 subjects. The study group consisted of subjects at high risk for silicosis extracted from a database of construction workers and miners. The control group consisted of heavy smokers from a lung cancer screening program. The groups were matched for age, sex and smoking behavior (research showed that the vast majority of construction workers with high silica exposure are also heavy smokers). Two radiologists visually, in a blinded fashion, scored how many micronodules were present in each scan. The authors found that in almost all scans micronodules were present (median 4 in both groups) and that in the control group the 95th percentile was 13 micronodules,

which they regarded as the upper limit of normal (ULN). Subjects with more than 13 micronodules should be considered at high risk for silicosis. In the group of construction workers 12 subjects (22%, significantly more than 5%), had >13 micronodules, while only two subjects had disseminated micronodules, one of them 144, compatible with the radiological definition of silicosis. Mets et al.[76] hypothesized that the presence of moderate numbers of micronodules is an early sign of silicosis.

Detecting micronodules is, however, a tedious task. Such small nodules are easy to miss because their size is often similar to cross-sections of small vessels. Even searching for larger nodules (>4mm), as should be done to detect possible early lung cancer, takes expert human observers between 10 and 15 minutes per scan. Detecting subjects at high risk for silicosis requires finding those patients which have more than 13 micronodules, which is even more challenging. The difficulty of this task is depicted in Fig.3.1. Counting micronodules seems therefore unfeasible in clinical practice unless the task can be automated. This was the purpose of the current study, and we had the data of Mets et al. and additional scans of subjects at high risk of developing silicosis at our disposal. To our knowledge, this is the first paper on automated detection of micronodules, even though the problem of detecting larger nodules has been widely studied[35,36,40,41,69,77].

3.2 Methods

Data

The database with CT scans of construction workers at high risk for silicosis, described in the publication by Mets et al.[76], was used to select cases for the development set. In that publication, the study group contained 54 out of the 159 subjects in the database: only the subjects which matched for age and smoking behavior with the control group were included. We selected 15 scans from the remaining subjects of the database as a development data set. On these 15 scans, a radiology resident annotated micronodules for training, 85 in total. Note that the radiology resident did not extensively search all 15 scans for micronodules, but created a subset of micronodules for training of the CAD algorithm.

For evaluation of the CAD system, all scans of the study group of 54 dust-exposed subjects were used. Two observers counted the number of micronodules in all 54 scans. The average of the scores of the two observers was used. The median number of micronodules on these scans was 4 with a range from 0 to 144.

All subjects received low-dose chest CT examination with either 16x0.75-mm collimation (Mx8000 IDT or Brilliance 16P; Philips Medical Systems, Cleveland, OH)

#	Description
1	Ratio: Max intensity $R1$/Max intensity $R2$
2	Ratio: Mean intensity $R1$/Mean intensity $R2$
3-4	Intensity and gradient magnitude of candidate voxel
5-8	Max, min, mean, std. dev. of intensities
9-12	Max, min, mean, std. dev. of CC
13-16	Max, min, mean, std. dev. of gradient vector magnitudes
17	Histogram entropy
18	Histogram mean position
19	Histogram std. dev.
20-21	Histogram mean height/nonzero mean height
22-23	Histogram max peak position/nonzero max peak position
24-25	Histogram max peak height/nonzero max peak height
26-30	Histogram quantile 5, 25, 50, 75, 95

Table 3.1: Features used for false positive reduction. All features are computed over region $R1$, except for features 1 and 2.

or 128x0.625-mm collimation (Brilliance iCT; Philips Medical Systems). Axial slices with a slice thickness of 1.0 mm were reconstructed with an interval of 0.7 mm using a moderately soft reconstruction kernel.

Candidate detection

The candidate detection stage consisted of four stages. First, the lung fields were automatically extracted using a method proposed by van Rikxoort et al.[52]. Second, the scan was cropped around the lung segmentation and resampled to an isotropic resolution of 0.7 mm^3. Third, all voxels above a threshold t were selected. Fourth, similar to Lee et al.[78], we performed template matching on selected voxels. A 3D Gaussian blob template with scale σ was used. Normalized cross-correlation coefficient (CCC) was used as a similarity measure. Finally, local maxima detection in a 26-neighborhood was applied to acquire the final candidate locations. During pilot experiments it was found that $t = -700$HU and $\sigma = 0.35$mm yielded optimal results on the training set.

False positive reduction

A set of 30 features was used here to further describe the candidates and to reduce false positive responses. They are listed in Table 3.1. Features were calculated from

a region around a candidate. Region 1 ($R1$) is a spherical region with a diameter of 3 mm with the candidate as center. This region has approximately the size of a micronodule. Region 2 ($R2$) is a spherical region around the candidate with a diameter of 8 mm, from which $R1$ has been excluded. The features can be subdivided into 4 classes: intensity features, gradient features, correlation features and histogram features.

Features 1,2,3 and 5-8 were defined with respect to the scan intensity. Since micronodules have a high density, they most likely have a higher attenuation on the scan than the surroundings, even when located close to blood vessels or other structures. Features 3 and 5-8 took into consideration the intensity in a small region around the candidate, while features 1 and 2 were also dependent on the statistics of the wider surroundings. Features 4 and 13-16 were obtained from the magnitude of the gradient vectors. The gradient vector was calculated using a Sobel kernel and measures the gradient in scan intensity and its direction. The magnitude of this vector is the total change of intensity in the region and is a scalar. This scalar was used in feature calculations. In a round object on a CT scan, the gradient vectors on the edge all radiate from the center. When the candidate represents a blood vessel, a common source of false positives, gradients will all point in the same direction, and their magnitudes will show a different pattern than that of a round object. Features 9-12 were calculated from the CCC of the candidate and its surroundings. In addition to the maximum CCC value, statistics of the surroundings are also of interest. Features 17-30 were calculated from the histogram of local scan intensity values in R1. The histograms have a bin width of 10 HU. Histogram features give a more detailed description of the intensity distribution around candidates, which can also be used to discriminate between micronodules and other objects.

Candidates were classified with a k-nearest neighbor classifier with k set to the square root of the number of training samples, which gave $k = 149$.

Optimizing the training data set using an active learning approach

During preliminary evaluation of the CAD algorithm, it became evident that there was a substantial amount of micronodules not annotated on our 15 training cases. Therefore, we used an active learning approach to improve the annotations on the training data set. In this approach, we use the CAD algorithm itself to update the annotations on the training data set. We start by performing a leave-one-out cross-validation experiment. Next, we perform FROC analysis to evaluate the performance of the system. Subsequently, we inspect the 100 most suspicious false positives and mark the false positives which are actually micronodules. These micron-

Iteration	# micronodules among 100 most suspicious false positives	Total # of micronodule annotations
1	63 / 100	85
2	37 / 100	148
3	9 / 100	185
4	0 / 100	194

Table 3.2: Results from optimizing the training data set using the active learning approach.

odules are then added to the reference standard on the training set and we repeat the same procedure again. Note that the training in the cross-validation now uses these new annotations and the performance of the system improves. This process is iterated until no micronodules are found among the 100 most suspicious false positives. Note that we are using the CAD algorithm to update the annotations on the training set and as a consequence, although leave-one-out cross-validation is applied, a biased estimate of the performance of our system on the training set will be obtained. An additional evaluation of our system will be performed on independent data, the study group used by Mets et al.[76].

3.3 Results

Four iterations of the active learning approach were performed to update the annotations of the training data set and the number of added micronodule annotations are shown in Table 3.2.

Leave-one-nodule-out cross validation was used to assess the performance of the micronodule CAD system on the final development set with 194 micronodule annotations. Free-response operating characteristic (FROC) analysis was applied to measure the performance of the CAD system. The FROC curve is depicted in Fig. 3.2 and shows a good performance with a sensitivity of 84% at an average of 8.4 false positives (FPs) per scan.

An experiment was conducted to test whether the CAD system can detect subjects with disseminated micronodules, compatible with radiological features of manifest silicosis. The micronodule counts by the radiologists suggested that 2 out of the 54 subjects in the study group had radiological patterns of manifest silicosis. We hypothesize that a micronodule CAD system set to operate at a high sensitivity is able to pick out cases with high micronodule counts. To do so, the CAD algorithm was set

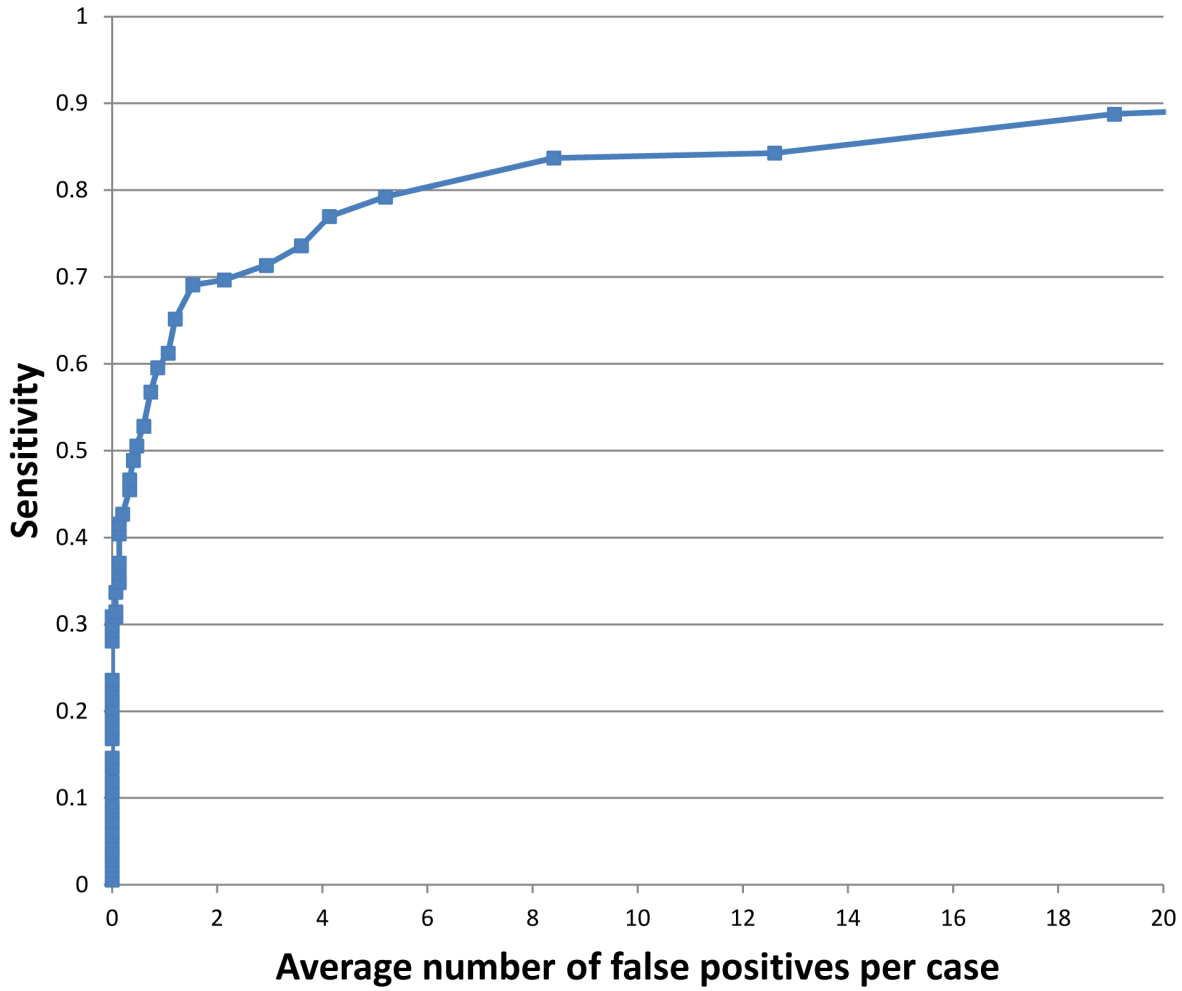


Figure 3.2: FROC curve of the micronodule CAD system, obtained using leave-one-nodule-out cross validation.

to operate at an operating point of 84% sensitivity and an average of 8.4 FPs/scan. The CAD system processed all 54 cases and we evaluated the potential of the CAD system running fully automatic, e.g. without visual checking of the CAD marks by a radiologist.

On the test set of 54 cases, the CAD system generated 637 CAD marks (11.8 CAD marks per scan), with a median of 3 (range 0-303). The two cases with a radiological pattern of manifest silicosis had the highest number of CAD marks, 303 and 45 CAD marks, respectively. If a threshold of 40 CAD marks per scan would be used, the two cases with manifest silicosis could be detected in this database without any false positive cases. Fig. 3.3 displays two axial sections of these two cases in which many micronodules are present.

We ordered the cases into four groups, (1) low risk: cases with less or equal than 4 micronodules, (2) intermediate risk: cases containing between 5 and 13 micronodules, (3) high risk: cases containing between 13 and 40 micronodules and (4) mani-

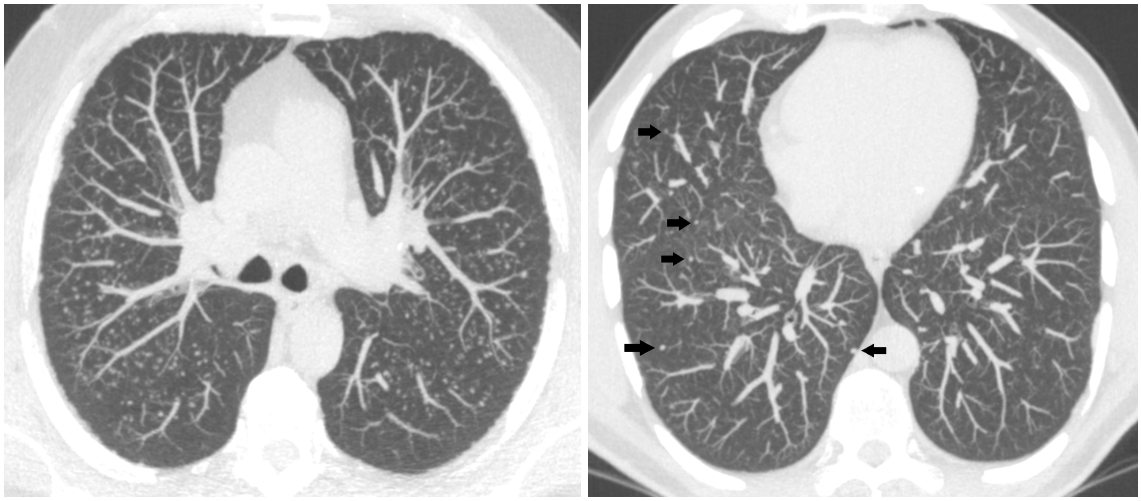


Figure 3.3: Two axial sections of the two scans with the most micronodules according to both CAD system and human observers. These two scans were the only ones compatible with radiological features of manifest silicosis. Both images are axial maximum intensity projections of 10 mm. In the right image, arrows indicate the detected micronodules (in the left image there were too many detections to indicate).

fest silicosis: cases with more than 40 micronodules. Note that cases in group 3 are at higher risk for silicosis and above the upper limit of normal (UPN) of the control group[76]. The confusion matrix which compares the reference standard set by the observers with the CAD system is displayed in Table. 3.3. The weighted kappa between the risk assessment by human observers and the CAD system was 0.76.

3.4 Discussion and Conclusion

In this study, we developed an automated detection system for micronodules that has good performance and can be used to automatically classify scans into low, intermediate, high risk, and manifest silicosis. The confusion matrix in Table. 3.3 shows that both cases of manifest silicosis were correctly detected without any false positive detections, and that only 2 out of 54 subjects were not assigned to either the correct risk group or a neighboring risk group. Manual annotation of micronodules by human observers is exhausting, labor intensive and prone to errors. A CAD tool may be able to give a more precise quantification which can be important for the purpose of monitoring disease progression. Therefore, an efficient CAD system for micronodules is of clinical significance. Future work includes experimenting with other classifiers, optimization and expansion of the feature set, and validation on a larger cohort.

An automated system for the detection and quantification of micronodules from

		CAD Prediction			
		Low Risk	Intermediate Risk	High Risk	Manifest Silicosis
Observers	Low Risk	28	4	2	0
	Intermediate Risk	1	9	2	0
	High Risk	0	0	6	0
	Manifest Silicosis	0	0	0	2

Table 3.3: Confusion matrix for the four risk categories.

thoracic CT scans has been presented. The system was validated by comparing with a counting by humans in a high risk screening cohort and shows excellent performance. This paves the way for automated risk assessment for silicosis, which could be useful to screen high risk subjects such as construction workers or miners, and could also be included in lung cancer screening programs.

Comparing CAD systems on a public database

4

C. Jacobs, E.M. van Rikxoort, K. Murphy, M. Prokop, C.M. Schaefer-Prokop and B. van Ginneken

Original title: Computer-aided detection of pulmonary nodules: a comparative study using the public LIDC/IDRI database

Published in: European Radiology, 2015

Abstract

Objectives To benchmark the performance of state-of-the-art computer-aided detection (CAD) of pulmonary nodules using the largest publicly available annotated CT database (LIDC/IDRI), and to show that CAD finds lesions not identified by the LIDCs four-fold double reading process.
Methods The LIDC/IDRI database contains 888 thoracic CT scans with a section thickness of 2.5 mm or lower. We report performance of two commercial and one academic CAD system. The influence of presence of contrast, section thickness, and reconstruction kernel on CAD performance was assessed. Four radiologists independently analyzed the false positive CAD marks of the best CAD system.
Results The updated commercial CAD system showed the best performance with a sensitivity of 82% at an average of 3.1 false positive detections per scan. Forty-five false positive CAD marks were scored as nodules by all four radiologists in our study.
Conclusions On the largest publicly available reference database for lung nodule detection in chest CT, the updated commercial CAD system locates the vast majority of pulmonary nodules at a low false positive rate. Potential for CAD is substantiated by the fact that it identifies pulmonary nodules that were not marked during the extensive four-fold LIDC annotation process.

4.1 Introduction

The last two decades have shown substantial research into computer-aided detection (CAD) of pulmonary nodules in thoracic computed tomography (CT) scans[22,23]. Although many academic and several commercial CAD algorithms have been developed, CAD for lung nodules is still not commonly used in daily clinical practice. Possible explanations for this are a lack of reimbursement, technical impediments to integration into PACS systems, but also low sensitivity and high false positive rates. The recent positive results of the NLST lung cancer screening trial[10] and the subsequent developments towards implementation of lung cancer screening in the United States[29,30] have renewed the interest into CAD for pulmonary nodules. If lung cancer screening will be implemented on a large scale, the burden on radiologists will be substantial and CAD could play an important role in reducing reading time and thereby improving cost-effectiveness[31,32].

Following the general demand for open and reproducible science, public databases have been established to facilitate objective measures of CAD performance, and to move CAD development to a next level[69,70,79]. In 2011, the complete LIDI/IDRI (Lung Image Database Consortium/Image Database Resource Initiative) database was released[70]. This dataset provides by far the largest public resource to assess the performance of algorithms for the detection of pulmonary nodules in thoracic CT scans. A large effort has gone into the collection of annotations on these cases, but CAD was not used to assist the readers[70].

In this paper, we apply two commercial and one state-of-the-art academic nodule detection system on the LIDC/IDRI database with the aim to set a first benchmark performance on the full database. To our knowledge, this is the first paper which reports the performance of CAD systems on the full LIDC/IDRI database. We performed an extensive analysis of the performance of the applied CAD systems and make our evaluation publicly available so that other CAD developers can compare with this benchmark. Furthermore, we hypothesize that CAD can find lesions which were not detected in the extensive LIDC annotation process consisting of a blinded and unblinded review by four radiologists. To investigate the latter, we evaluated the false positives of the best CAD system using a similar reading protocol as had been used in LIDC.

Manufacturer	Model name	Number
GE MEDICAL SYSTEMS	LightSpeed 16	197
GE MEDICAL SYSTEMS	LightSpeed Ultra	162
GE MEDICAL SYSTEMS	LightSpeed QX/i	97
GE MEDICAL SYSTEMS	LightSpeed Pro 16	79
GE MEDICAL SYSTEMS	LightSpeed VCT	61
GE MEDICAL SYSTEMS	LightSpeed Plus	56
GE MEDICAL SYSTEMS	LightSpeed Power	10
Philips	Brilliance 16P	54
Philips	Brilliance 64	49
Philips	Brilliance 40	9
Philips	Brilliance 16	5
SIEMENS	Sensation 16	95
SIEMENS	Sensation 64	5
SIEMENS	Definition	3
SIEMENS	Emotion 6	1
TOSHIBA	Acquilion	5
Total		888

Table 4.1: Manufacturer and scanner model distribution of the 888 CT scans in our dataset.

4.2 Materials and Methods

Data

This study used the LIDC/IDRI data set[70], consisting of 1,018 helical thoracic CT scans collected retrospectively from seven academic centers. Nine cases with inconsistent slice spacing or missing slices were excluded. In addition, 121 CT scans which had a section thickness of 3 mm and higher were excluded since thick section data is not optimal for CAD analysis. This resulted in 888 CT cases available for evaluation. In Tables 4.1, 4.2, and 4.3, the characteristics of the input data are shown.

LIDC/IDRI image annotation

The LIDC/IDRI employed a two-phase image annotation process[70]. In the first phase (the blind phase), four radiologists independently reviewed all cases. In the second phase (the unblinded phase), all annotations of the other three radiologists were made available and each radiologist independently reviewed their marks along with the anonymized marks of their colleagues. Findings were annotated and cate-

Section thickness [mm]	Number
0.6	7
0.75	30
0.9	2
1	58
1.25	343
1.5	5
2	123
2.5	320
Total	888

Table 4.2: Section thickness distribution of the 888 CT scans in our dataset.

Manufacturer and reconstruction kernel	Type	Number
GE MEDICAL SYSTEMS - BONE	Enhancing	220
GE MEDICAL SYSTEMS - LUNG	Overenhancing	70
GE MEDICAL SYSTEMS - STANDARD	Standard	372
Philips - B	Standard	21
Philips - C	Enhancing	7
Philips - D	Overenhancing	45
SIEMENS - B20s	Soft	1
SIEMENS - B30f	Standard	102
SIEMENS - B31f	Standard	1
SIEMENS - B45f	Enhancing	30
SIEMENS - B50f	Enhancing	2
SIEMENS - B70f	Overenhancing	12
TOSHIBA - FC03	Standard	2
TOSHIBA - FC10	Soft	3
Total		888

Table 4.3: Distribution of the reconstruction kernels used for the 888 CT scans in our dataset.

gorized into *nodule≥3mm*, *nodule<3mm*, or *non-nodule*. *non-nodule* marks were used to indicate abnormalities in the scan which were not considered a nodule. Using this two-phase process, the LIDC investigators aimed to identify as completely as possible all lung nodules, without forcing consensus among the readers. More details about the annotation process can be found in Armato et al.[70]. An XML file with the annotations is publicly available for every case.

Nodule selection and purpose

In this study, we included all annotations available in the XML files for the 888 scans. The focus of this study lied on the *nodule≥3mm* group. As a result of the LIDC/IDRI image annotation process, each *nodule≥3mm* had been annotated by one, two, three or four radiologists. In total, the data set of this study included 777 locations which were marked as *nodule≥3mm* by all four radiologists. The 777 *nodule≥3mm* annotations marked by all four radiologists can be categorized by size as follows: 22 nodules <4 mm, 228 nodules 4-6 mm, 199 nodules 6-8 mm, and 328 nodules >8 mm. The number of nodules per scan ranged between 1 and 8.

The purpose of this study was twofold. First, we aimed to assess the performance of three state-of-the-art nodule CAD systems. Secondly, we performed an observer experiment to investigate whether CAD can find additional lesions, missed during the extensive LIDC annotation process.

CAD systems

Three CAD systems were used: a commercial CAD system *Visia* (MeVis Medical Solutions AG, Bremen, Germany), a commercial prototype CAD system *Herakles* (MeVis Medical Solutions AG, Bremen, Germany), and an academic nodule CAD system *ISICAD* (Utrecht Medical Center, Utrecht, the Netherlands)[36]. *ISICAD* was the leading academic CAD system in the ANODE09 nodule detection challenge[69]. For all three CAD systems, a list of candidate marks per CT scan was obtained. Each CAD candidate is described by a 3D location. Additionally, *Herakles* and *ISICAD* also provide a CAD score per CAD candidate. The CAD score is the output of the internal classification scheme of the CAD system and is a measure of the likelihood that a candidate is a pulmonary nodule. An internal threshold on the CAD scores determines which candidates are active CAD marks and hence will be shown to the user, and which candidates are not shown. Since different thresholds can be applied on the CAD score, a CAD system can have multiple operating points. A low threshold generates more CAD marks, thereby typically increasing sensitivity at the cost of more false positive detections. A high threshold will generate less false positives but

may reduce the sensitivity of a CAD system. For all three CAD systems, one fixed operating point is internally set which we will refer to as the system operating point.

Evaluation

The performance of the CAD systems was analyzed on the set of 777 nodules annotated by 4/4 radiologists as a *nodule≥3mm*. We employed free-response operating characteristic (FROC) analysis[80] where detection sensitivity is plotted against the average number of false positive detections per scan. Confidence intervals were estimated using bootstrapping with 5,000 iterations[81]. If a CAD system marked locations which were annotated by three or fewer radiologists as *nodule≥3mm*, as *nodule<3mm*, and as *non-nodule*s, these CAD marks were counted as false positives. For *Visia*, no CAD scores were available for the CAD candidates. Consequently, only one operating point and not a full FROC curve could be generated for *Visia*.

To gain more insight into which type of nodules were missed by CAD, we looked at the characteristics, as scored by the LIDC readers for all *nodule≥3mm* findings, of the false negatives. We defined subsolid nodules as nodules for which the majority of the radiologists gave a texture score smaller than 5 (1=ground-glass/non-solid, 3=part-solid, 5=solid). Subtle nodules were defined as nodules for which the majority of the radiologists gave a subtlety score smaller or equal than 3 (1=extremely subtle, 5=obvious).

To assess the robustness of the CAD algorithms, we also evaluated the CAD results on different subsets of the data. The LIDC-IDRI data set is a heterogeneous set of CT scans and CAD algorithms could conceivably exhibit different performance on different types of data. We analyzed the following factors: (1) presence of contrast material, i.e. non-contrast versus contrast enhanced scans, (2) section thickness, i.e. cases with section thickness $<$2mm versus section thickness $\geq$2mm, and (3) reconstruction kernel, i.e. soft/standard versus enhancing/overenhancing kernels.

Observer study

In order to evaluate whether CAD can find lesions missed during the extensive annotation process of the LIDC/IDRI database, we considered the CAD marks of the best CAD algorithm which were counted as false positives at its system operating point. Two conditions were differentiated: the location of the CAD mark had in fact been marked in the LIDC annotation process, but not by all four readers as *nodule≥3mm* as warranted for being counted as a true positive. The second condition comprised those CAD marks that had no corresponding LIDC marks at all. The CAD marks corresponding to the first condition can be subdivided according to the

LIDC readings. The latter CAD marks were independently inspected by four chest radiologists, since these are potentially nodules overlooked by all four LIDC readers. Thus we mimic the original LIDC annotation process as though CAD had been included as another independent reader in the first phase of the image annotation process. CAD marks were categorized as *nodule≥3mm*, *nodule<3mm*, *non-nodules*, or false positive. Electronic measurement tools were available to measure size. To reduce the workload for the radiologists, a research scientist (5 years experience in nodule CAD research) first removed the marks which were obviously not a nodule. CAD marks which were marked as *nodule≥3mm* by all four radiologists in our study were independently evaluated by an experienced radiologist that scored subtlety, location, type, and attachment to other structures. Subtlety was scored on a five-point scale (1=extremely subtle, 5=obvious).

4.3 Results

Comparative CAD performance

The performance of the three CAD systems is depicted in Fig 4.1. From the FROC curves it is evident that *Herakles* performed best. The system performances were significantly different ($p<0.001$). At its system operating point, *Herakles* reached a sensitivity of 82% at an average of 3.1 false positives per scan for nodules all four LIDC readers had agreed on.

We evaluated the characteristics of the 141 false negative nodules. 42 (30%) false negatives were subsolid nodules. The size distribution of the missed nodules was as follows: 5 nodules ¡4mm, 53 nodules 4-6 mm, 31 nodules 6-8 mm, and 52 nodules >8 mm. Thus, a large portion of the missed nodules were smaller than 6 mm, but still a substantial number of missed nodules, 52 (37%), were larger than 8 mm. Finally, we found that 33 (23%) of the missed nodules were subtle. Fig. 4.2 shows eight randomly chosen missed nodules.

The performance of the three CAD systems on the different subsets is depicted in Fig 4.3. This figure shows that the performance of *ISICAD* and *Visia* was influenced by different data sources. *ISICAD* shows the largest performance difference between soft/standard versus enhancing/overenhancing reconstruction kernels. *Herakles* showed the most stable and robust performance for all different data sources and consistently outperformed the other two CAD systems.

We categorized the CAD marks of *Herakles* which were counted as false positives at its system operating point. In total, there were 2,720 false positive CAD marks in the 888 cases (Table 4). The majority of the CAD marks, 1,612 out of 2,720 (59%), had

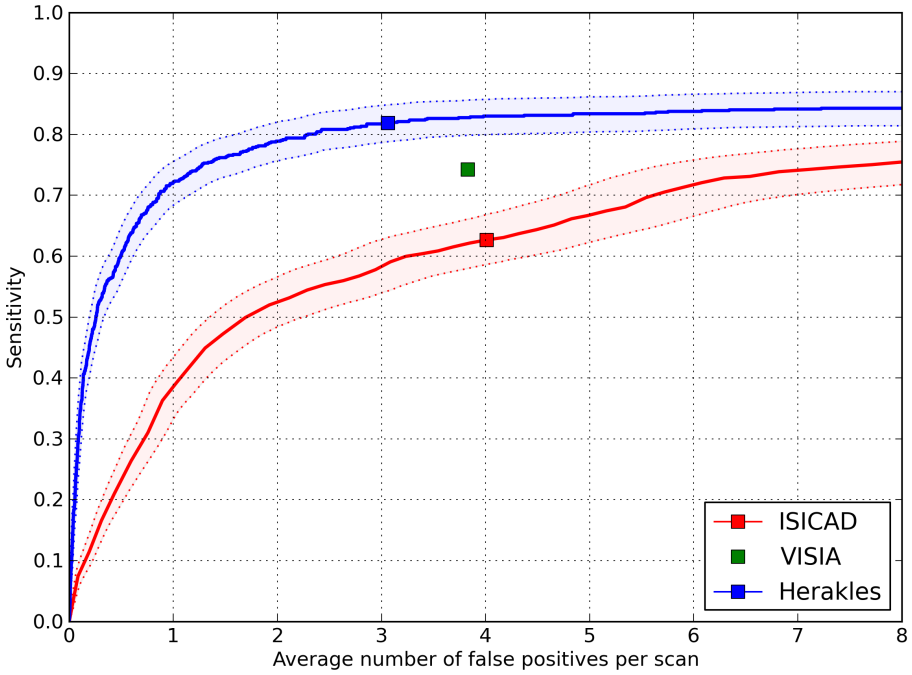


Figure 4.1: FROC curves for all three CAD systems on the full database of 888 CT scans containing 777 nodules for which all four radiologists classified it as *nodule≥3mm*. The points on the curves indicate the system operating points of the three CAD systems. For *Visia*, no continuous FROC curve but only a single operating point can be provided since the CAD scores of the CAD marks are not available. Shaded areas around the curve indicate 95% confidence intervals.

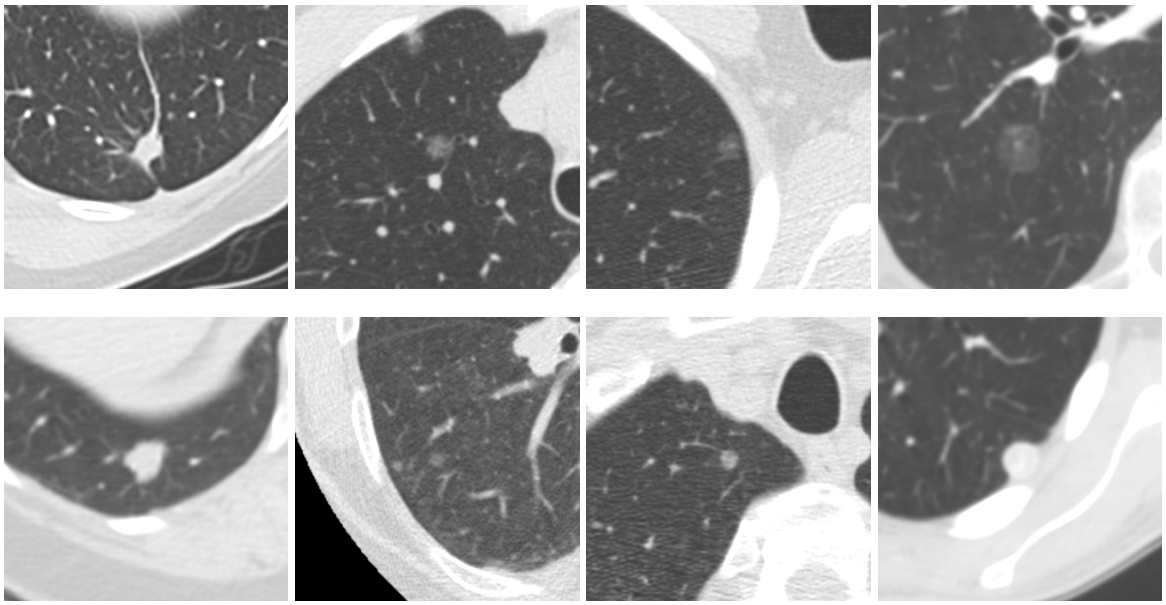
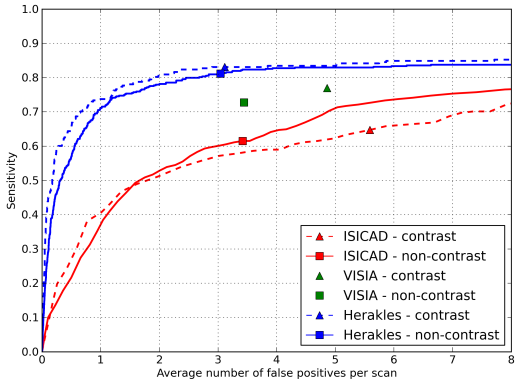
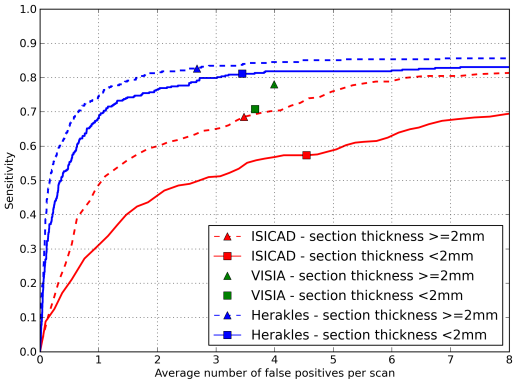


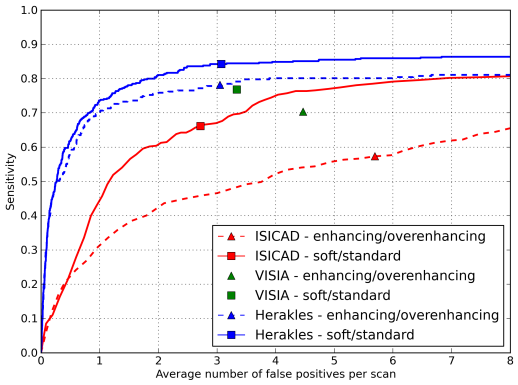
Figure 4.2: Eight randomly chosen examples of false negatives of *Herakles*. Each image shows a transverse field of view of 60 x 60 mm in which the nodule is centered. Note that many missed nodules are subsolid.



(a) contrast vs non-contrast



(b) slice thickness $<$2mm vs $\geq$2 mm



(c) soft/standard vs enhancing/overhancing reconstruction kernels

Figure 4.3: FROC curves for all three CAD systems on (A) contrast scans (n=242) versus non-contrast scans (n=646), (B) scans with a section thickness $<$2 mm (n=445) versus scans with a section thickness $\geq$2 mm (n=443), and (C) scans with a soft or standard reconstruction kernel (n=502) versus scans with an enhancing or overenhancing reconstruction kernel (n=386). The reference set of nodules consists of nodules for which all four radiologists classified it as *nodule*$\geq$*3mm*. The points on the curves indicate the system operating points of the three systems. For *Visia*, no continuous FROC curve but only a single operating point can be provided since the CAD scores of the CAD marks are not available.

at least one corresponding mark from the LIDC readers. These CAD marks can be further categorized into marks on annotations marked as *nodule*$\geq$*3mm* by 3 out of 4 radiologists, 2 out of 4 radiologists, 1 out of 4 radiologists, annotations marked as *nodule*$<$*3mm* by at least 1 radiologist (and hence no *nodule*$\geq$*3mm* annotations), and finally annotations marked as *non-nodule* by at least 1 out of 4 radiologists (and hence no *nodule*$\geq$*3mm* or *nodule*$<$*3mm* annotations). Table 4.4 shows how the CAD marks were further split out into these categories. The remaining 1,108 false positive CAD

marks had no corresponding mark from the LIDC readers.

Category	Number
nodule≥3mm - 3/4	254
nodule≥3mm - 2/4	208
nodule≥3mm - 1/4	219
nodule<3mm	423
non-nodule	508
no corresponding annotation	1108
Total	2720

Table 4.4: Overview of the categories in which the false positives of *Herakles* at the system operating point can be divided. In this analysis, we first check for corresponding *nodule≥3mm* annotations, then we check for corresponding *nodule<3mm* annotations, and finally we check for corresponding *non-nodule* annotations. This means that in the top row where 3 out of 4 radiologists annotated the location as *nodule≥3mm*, the fourth radiologist may have marked the location as *nodule<3mm*, *non-nodule*, or did not mark it at all. In the *nodule<3mm* category, all false positives whose location was marked as *nodule<3mm* by at least one radiologist were placed (and hence no radiologist marked it as *nodule≥3mm*). The non-nodule category contains all false positives whose location was marked as *non-nodule* by at least one radiologist (and hence no radiologist marked the location as *nodule≥3mm* or *nodule<3mm*). False positives for which no corresponding annotation was found were assigned to the last category.

Observer study results

In our observer experiment, we focused on these 1,108 false positive CAD marks of *Herakles* which had no corresponding mark from any of the LIDC readers. These are locations which were potentially overlooked by all four LIDC readers. After CAD marks which were obviously not a nodule had been removed by the research scientist, 269 CAD marks were left for analysis by the four radiologists. Common sources of false positive detections removed by the research scientist included fissure thickening at the chest wall, vessel bifurcations and (micro-)atelectasis. Table 4.5 depicts how each of the observers scored these 269 CAD marks. In total, 45 CAD marks were considered to be a *nodule≥3mm* by all four radiologists; 177 CAD marks were considered to be a *nodule≥3mm* by at least one of the radiologists. The size distribution of the 45 CAD marks was as follows: 9 nodules <4 mm, 27 nodules 4-6 mm, 7 nodules 6-8 mm, and 2 nodules >8 mm. Subtlety was scored lower or equal than 3 for 32 (71%) nodules. Location was scored as central for 11 nodules,

Type	Observer 1	Observer 2	Observer 3	Observer 4
nodule≥3mm	119	97	84	153
nodule<3mm	125	141	136	50
non-nodule	20	20	46	41
false positive	5	11	3	25
Total	269	269	269	269

Table 4.5: Results of the observer experiment. The distribution of the scores of all observers is tabulated.

peripheral for 11 nodules, and in-between for 23 nodules. Nodule type was scored as follows: 32 solid, 2 ground-glass, 1 part-solid, and 10 calcified. Vascular, pleural or fissural attachment was found for 18 (40%) nodules. Fig. 4.4 shows eight randomly chosen examples of CAD marks which were considered a nodule3mm by all four radiologists and were scored as solid. In addition, 169 marks were considered a *nodule≥3mm* or a *nodule<3mm* by all four radiologists; 250 marks were considered a *nodule≥3mm* or a *nodule<3mm* by at least one of the radiologists. Thus, following the reference of the 4-reader agreement and adding these 45 CAD marks to the set of nodules, the updated performance of *Herakles* at its system operating point would reach a sensitivity of 83% at an average of 3.0 false positive detections per scan. In this FROC analysis, CAD marks on locations marked as *nodule≥3mm* by 3 out of 4 radiologists, 2 out of 4 radiologist, 1 out of 4 radiologists, or as *nodule<3mm* by at least 1 radiologist were counted as false positives. Evidently, one could argue whether CAD marks on these locations should be counted as false positives or not. If CAD marks on these locations were not to be counted as false positives but ignored in the FROC analysis, a performance of 83% sensitivity at an average of only 1.0 false positives per scan would be reached.

4.4 Discussion

Though clear definitions are available of what represents a pulmonary nodule (Fleischner Glossary[12]), the literature lists a number of publications demonstrating the lack of observer agreement of what indeed represents a pulmonary nodule[17,19,82]. Not surprisingly this effect is larger for small lesions[17]. This lack of an absolute standard of truth makes benchmarking of CAD systems very difficult. Therefore we decided to use the largest publicly available database of CT annotated pulmonary nodules. An elaborate double reading process involving 4 radiologists had been undertaken to define various levels of evidence for the presence of nodules to avoid the

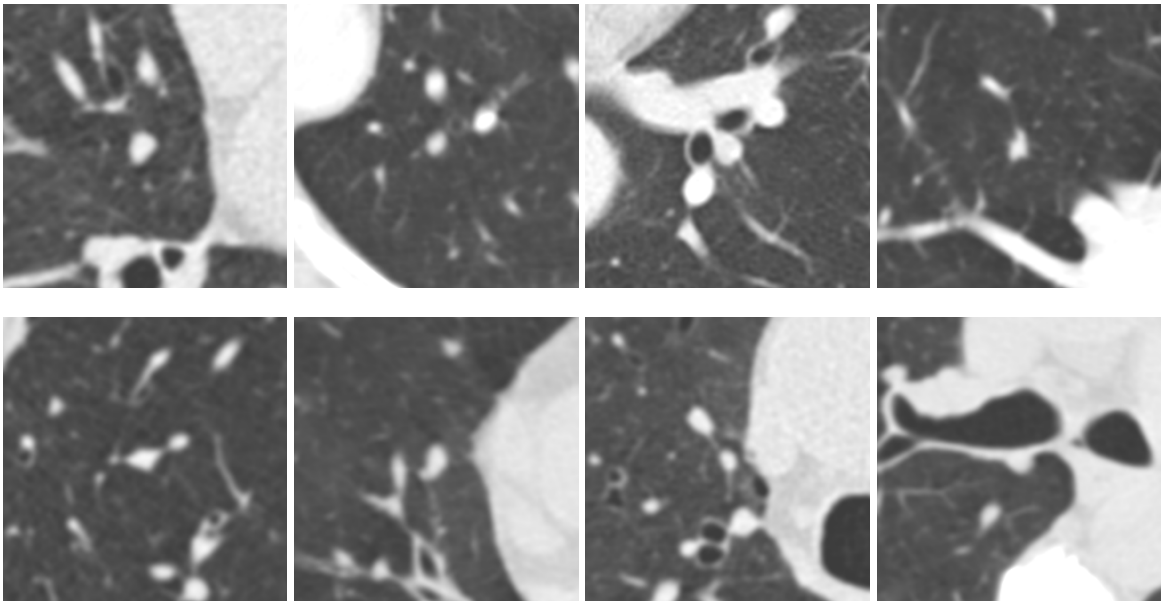


Figure 4.4: Eight randomly chosen examples of solid nodule annotations marked as *nodule≥3mm* by all four readers in our observer experiment. These nodules were not annotated by any of the original LIDC readers. Each image shows a transverse field of view of 60 x 60 mm in which the nodule is centered.

need for a consensus statement. In our study we used the extensive annotation information of the LIDC/IDRI database to benchmark the performance of state-of-the-art nodule CAD systems. To our knowledge this is the first study that uses the full LIDC database and secondly accepts the fact that there is no absolute standard of truth for the presence of pulmonary nodules in the absence of pathological correlation.

Our study showed substantial performance differences between the three CAD systems, with the commercial prototype *Herakles* demonstrating the best performance. At its system operating point, *Herakles* detected 82% of all *nodule≥3mm* findings marked by all four LIDC readers at an average of 3.1 false positives per scan. If marks on the other LIDC annotations were ignored in the analysis, a sensitivity of 83% at an average of only 1.0 false positives was reached.

The best CAD system still misses a subset of the nodules (18% of the 777 nodules). We observed that a substantial part of the missed nodules (30%) were subsolid nodules, which are more rare and have a markedly different internal character than solid nodules. Therefore, integrating a dedicated subsolid nodule detection scheme[77] in a complete CAD solution for pulmonary nodules may prove helpful to improve overall CAD performance.

Both *Visia* and *ISICAD* showed substantial performance differences on different subsets of the data, but *Herakles* achieved a more robust performance. The performance of *ISICAD* dropped substantially on data with enhancing or overenhancing

reconstruction kernels. This may be attributed to the fact that *ISICAD* was developed and trained exclusively with data from the Dutch-Belgian NELSON lung cancer screening trial, which consists of homogeneous thin-slice data reconstructed with a soft/standard reconstruction kernel[9]. This indicates that although *ISICAD* was the leading CAD system for the data used in the ANODE challenge[69], which consisted only of data obtained from the NELSON trial, its performance drops when applied to data of other sources. Therefore, the heterogeneity of a reference database is an important aspect for a reliable CAD evaluation and an advantage of the LIDC/IDRI database.

Although a blinded and unblinded review of all images had been performed by the LIDC investigators, we showed that CAD can find lesions missed by the original LIDC readers. We found 45 nodules which were accepted as a *nodule≥3mm* by all four radiologists involved in our observer study. Previous studies have already shown that CAD can find lesions missed by multiple readers[77,83]. One possible reason why the LIDC readers missed nodules may be that the LIDC readers only inspected transverse sections[70]. Characteristic features of the 45 nodules not included in the LIDC/IDRI database but seen by CAD were subtle conspicuity, small size (<6 mm) and attachment to pleura or vasculature.

Since an extensive evaluation on a large reference database is essential to move CAD to the next level, we have published our results on a public website (`http://luna.grand-challenge.org`) which allows other CAD researchers to upload results of their CAD systems for which the same FROC curves as presented in Fig. 4.1 and Fig. 4.3 will be computed and published on the website. The annotation files of the reference standard and the extra annotations by the human readers in our observer study are available for download. By making the extra annotations available to other researchers, this study contributes to an improved reference standard for the LIDC/IDRI database and we hope future CAD studies will use the improved reference standard.

We primarily evaluated the performance of CAD on nodules for which all four radiologists agreed that it was a *nodule≥3mm*. Previous publications have also focused on the nodules detected by three, two, or one out of four radiologists[84,85]. For using CAD in a screening setting, a high sensitivity even at the expense of specificity is desirable to find all potential cancerous nodules. High false positive rates on the other hand increase the workload to radiologists and potentially increase unnecessary follow-up. We therefore report the sensitivity using the highest level of evidence (4 out of 4 readers) and considered the lower levels of agreement for quantifying the false positive rates. For future CAD reference databases, a large database of CT images including follow-up CT and histopathological correlation would be helpful to

remove subjectivity from the reference standard, and to verify whether CAD detects the clinically relevant nodules.

In conclusion, we found that on the largest publicly available database of annotated chest CT scans for lung nodule detection, *Herakles* detects the vast majority of pulmonary nodules at a low false positive rate. The results show that the new prototype outperforms the other two CAD systems and is robust to different acquisition factors, such as presence of contrast, section thickness, and reconstruction kernel. Our observer experiment showed that *Herakles* was able to pulmonary nodules that had been missed by the extensive LIDC annotation process. Given the growing interest and need for CAD in the context of screening it can be expected that new CAD algorithms will be presented in the near future. Our results are publicly available and other CAD researchers may compare the performance of their CAD algorithm to the results reported here, utilizing the LIDC/IDRI database for benchmarking of available CAD systems.

Acknowledgments

The authors acknowledge the National Cancer Institute and the Foundation for the National Institutes of Health, and their critical role in the creation of the free publicly available LIDC/IDRI Database used in this study. This work was supported by a research grant from MeVis Medical Solutions AG, Bremen, Germany and by a research grant from the Netherlands Organisation for Scientific Research, project number 639.023.207.

Solid, Part-Solid, or Non-solid?

5

C. Jacobs, E.M. van Rikxoort, E.T. Scholten, P.A. de Jong, M. Prokop, C. Schaefer-Prokop and B. van Ginneken

Original title: Solid, Part-Solid, or Non-solid? Classification of Pulmonary Nodules in Low-Dose Chest Computed Tomography by a Computer-Aided Diagnosis System

Published in: Investigative Radiology 50:168-173, 2015

Abstract

Objectives: The purpose of this study was to develop and validate a computer-aided diagnosis (CAD) tool for automatic classification of pulmonary nodules seen on low-dose computed tomography into solid, part-solid, and non-solid.

Materials and Methods: Study lesions were randomly selected from 2 sites participating in the Dutch-Belgian NELSON lung cancer screening trial. On the basis of the annotations made by the screening radiologists, 50 part-solid and 50 non-solid pulmonary nodules with a diameter between 5 and 30 mm were randomly selected from the 2 sites. For each unique nodule, 1 low-dose chest computed tomographic scan was randomly selected, in which the nodule was visible. In addition, 50 solid nodules in the same size range were randomly selected. A completely automatic 3-dimensional segmentation-based classification system was developed, which analyzes the pulmonary nodule, extracting intensity-, texture-, and segmentation-based features to perform a statistical classification. In addition to the nodule classification by the screening radiologists, an independent rating of all nodules by 3 experienced thoracic radiologists was performed. Performance of CAD was evaluated by comparing the agreement between CAD and human experts and among human experts using the Cohen κ statistics.

Results: Pairwise agreement for the differentiation between solid, part-solid, and non-solid nodules between CAD and each of the human experts had a κ range between 0.54 and 0.72. The interobserver agreement among the human experts was in the same range (κ range, 0.56-0.81).

Conclusions: A novel automated classification tool for pulmonary nodules achieved good agreement with the human experts, yielding values in the same range as the interobserver agreement. Computer-aided diagnosis may aid radiologists in selecting the appropriate workup for pulmonary nodules.

5.1 Introduction

At present, the 5-year survival rate of patients diagnosed with lung cancer is very low with only 16%. This is mainly caused by the fact that only 15% of all diagnosed lung cancers are detected at an early stage[86]. Results by the National Lung Screening Trial, showing a 20% reduction in lung cancer mortality in a study group that received 3 annual low-dose computed tomographic (LDCT) scans[10], has fueled the debate on lung cancer screening using LDCT and generally raised the importance of appropriate management of incidentally found pulmonary nodules. At this early tumor stage, differentiation of benign versus malignant nodules largely depends on CT morphological criteria and their changes over time. This has been taken into account by the recommendations of the Fleischner Society that propose different management strategies for solid, part-solid, and non-solid nodules based on their different biological behavior[50,51,87]. These guidelines reflect the current evidence that shows that part-solid and non-solid nodules have a higher malignancy rate than solid nodules do, especially part-solid nodules[20]. Moreover, a recent publication, which presented an externally validated model to predict the malignancy likelihood of pulmonary nodules on baseline screening CT scans, showed that the nodule type, using the same 3 subgroups, was one of the most important predictors for malignancy[88]. This underlines the importance of an accurate assessment of the nodule type.

Studies investigating interobserver and intraobserver agreement between radiologists for the classification of pulmonary nodules are scarce: Yildirim et al[89] found a good interobserver agreement of experienced readers for differentiating solid from subsolid nodules (κ = 0.619 and κ = 0.654); however, in this study, no further differentiation was made for subsolid nodules into part-solid and non-solid lesions. Another study by van Riel et al[90] reported low interobserver agreement (κ = 0.33) and moderate intraobserver agreement (κ = 0.54) for readers of varying experience for classifying screening-detected solid, part-solid, and non-solid nodules on LDCT scans.

Computer-aided diagnosis (CAD) might be helpful for classification of pulmonary nodules. It can possibly reduce interobserver and intraobserver variability between human readers, especially between experienced and less-experienced readers. In addition, computer analysis algorithms aiming to predict the malignancy likelihood of pulmonary nodules most certainly would benefit from an automated classification. The purpose of this study was therefore to develop and validate a CAD tool for automatic classification of pulmonary nodules seen with LDCT into solid, part-solid, or non-solid.

5.2 Materials and Methods

Data

Data for this study were extracted from all CT scans from 2 sites of the Dutch-Belgian NELSON lung cancer screening trial[9]. The NELSON trial was approved by the Dutch Ministry of Health and the institutional review boards of the participating centers. Written informed consent was obtained from all participants. The original approval and informed consent included the ability to use data for future research, including retrospective studies. The NELSON database contains all nodule annotations with respect to localization as well as classification as solid, part-solid, or non-solid as determined by the local radiologists involved in the screening trial. The NELSON trial used a double reading paradigm, and the level of experience ranged between none to more than 20 years of experience reading thoracic CT scans for the first readers; both second readers had 6 years of experience[9].

Nodule selection

We randomly selected 50 nodules marked as part-solid and 50 nodules marked as non-solid in the range between 5 and 30 mm from the 2 sites of the screening trial. Most nodules were visible on multiple CT scans. For each unique nodule, we randomly selected 1 scan on which the nodule was visible. In addition, 50 nodules marked as solid in the same size range as the subsolid lesions were randomly selected from the NELSON database. None of the lesions was included twice; however, 1 CT scan could contain more than 1 lesion. Histopathology was not available for the majority of these nodules. The size of the nodules ranged from 6.0 mm to 28.3 mm in diameter with a median of 11.5 mm. In total, the CT data consisted of 126 LDCT scans originating from 117 subjects.

CT Imaging Protocol

All CT data had been acquired at full inspiration using the low-dose technique (16 x 0.75 mm; 120-140 kV(p), 30 mAs) using a 16-detector row CT scanner, either an MX8000 IDT/Brilliance 16 (Philips Medical Systems, Cleveland, OH) using a moderately soft reconstruction kernel (B; Philips Medical Systems) or a SOMATOM Sensation 16 (Siemens Medical Solutions, Erlangen, Germany) using a B30f kernel. The data set included 65 CT scans performed at 120 kV(p); the remaining 61 CT cases were performed at 140 kV(p). All reconstructions were performed using a 512 x 512 reconstruction matrix set around the widest dimension of the thorax. The in-plane

voxel size varied between 0.54 and 0.81 mm, and section thickness was 1 mm with an increment of 0.7 mm.

Automated Classification System

The automated classification CAD uses a previously published 3-dimensional nodule segmentation algorithm that combines region growing with a dedicated sequence of morphological operations to remove adjacent structures such as vessels and the pleural wall from the nodule[57]. The only human input is a user-defined seed point, which is expected to be placed near the center of the pulmonary nodule. On the basis of the seed point, a 3-dimensional volume of interest around the seed point is defined. Then, 2 different segmentation processes are started by using different parameter settings for the segmentation algorithm. The 2 different parameter settings are aimed (1) to segment the complete nodule, which may contain ground-glass and/or solid components, and (2) to segment the solid component of the nodule. A segmentation solely of the ground-glass components of the nodule is subsequently obtained by computing the relative complement of these 2 segmentations. For the 2 segmentation processes, different lower thresholds in Hounsfield units (HUs) were defined: 750 HU for the complete nodule and 450 HU for the solid component, respectively. The threshold for solid nodules and its effect on the final segmentation have been reported previously.11 The lower threshold for the ground-glass components has been adopted from a phantom study[91]. In case of a part-solid lesion, the seed point is expected to be placed inside the solid core of the lesion. Because the segmentation algorithm incorporates a step to remove attached vasculature[57], the solid core can be successfully segmented, even if vasculature is running through the ground-glass component of the lesion. An example of the different segmentations is given in Figure 5.1.

On the basis of these 3 segmentations, numerical characteristics (features) are extracted, taking the full 3-dimensional information into account. Volume, mass, and average density are calculated for each of the 3 segmentations. In prior experiments, however, we had found that a simple classifier based on only these features works insufficiently because it does not take into account the internal density distribution of the lesion. Therefore, a histogram analysis of the voxel densities inside the complete nodule segmentation is performed using a bin size of 10 HU. On the basis of the histogram, the following features are computed to describe the intensity distribution: entropy, standard deviation, mean height of all bins, density and height of bin with most voxels, and quantiles at 5%, 25%, 75%, and 95%. We assess the internal structure of the lesion on CT by extracting localized texture features, which we ob-

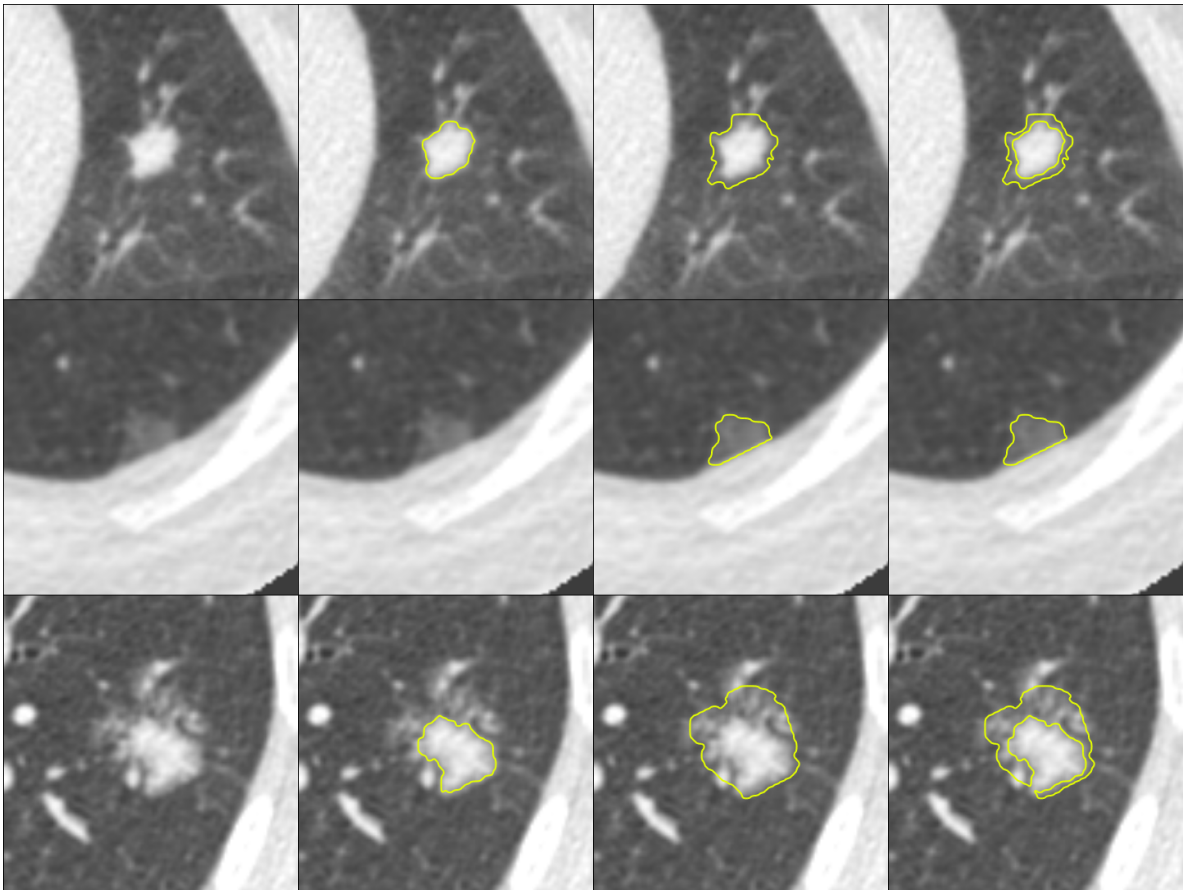


Figure 5.1: Illustrations of the different, possibly empty segmentations for a solid lesion (top row), non-solid lesion (middle row), and part-solid lesion (bottom row). Each subimage displays a transversal field of view of 6 x 6 cm. The left column shows the original image without segmentation results; the left middle column, solid component segmentation; the right middle column, complete nodule segmentation; and the right column, ground-glass component segmentation.

tain using local binary patterns[60]. Figure 5.2 illustrates 3 examples of nodules with a heterogeneous internal structure.

Finally, a regression k-nearest-neighbor classifier[65], which is a supervised statistical classifier, is applied to classify nodules into the 3 categories. The parameter k of the classifier is set to 12, the square root of the number of samples. Because the transition from a non-solid nodule to a solid nodule is a gradual process, a regression classifier was adopted. For training of the regression classifier, the response variable y was set to -1, 0, and 1 for non-solid, part-solid, and solid nodules, respectively.

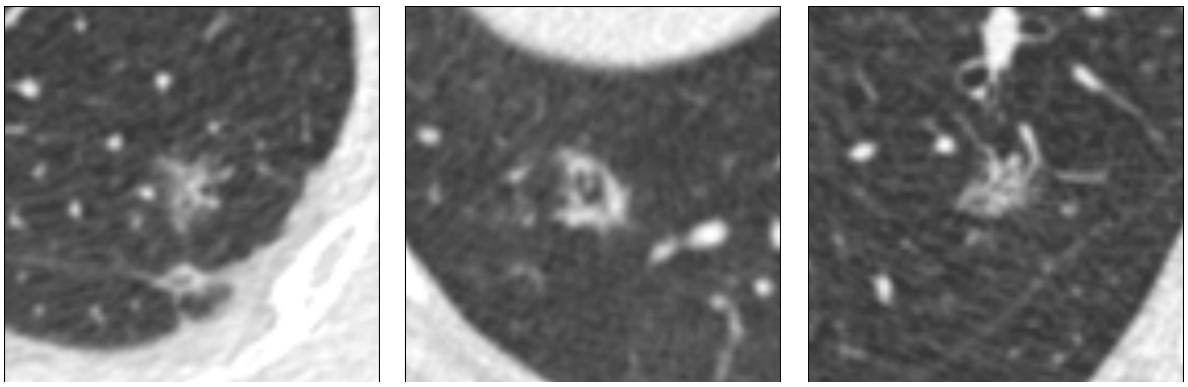


Figure 5.2: Illustrations of 3 nodules with a heterogeneous internal structure. Each subimage displays a transverse field of view of 6 x 6 cm. A clear solid core is difficult to appreciate, but all 3 lesions were categorized as part-solid by expert radiologists.

Observer study

An observer study was carried out involving 3 expert chest radiologists (all with >15 years experience in reading chest CTs) to assess the human interobserver agreement for the task at hand. The 3 independent expert radiologists will be referred to as reader 1, reader 2, and reader 3 in the rest of the article; the original rating by the screening radiologists will be referred to as reader 4. Readers 1, 2, and 3 independently classified all nodules as solid, part-solid, or non-solid having the full 3-dimensional thin section data set available, with display tools such as scrolling, windowing, magnification, and maximum intensity projections. In addition, they could indicate whether they considered a finding not to be a nodule. The latter judgment led to the removal of 12 lesions of the subsequent data analysis because they were deemed not to be nodules by all thoracic radiologists.

CAD Training Data Set

For the purpose of training the statistical classifier in the CAD algorithm, a training data set is required with a single classification per nodule. We decided to use the consensus opinion of 2 radiologists on each nodule. Because we could not consult the screening radiologists who were involved in the trial (reader 4), the opinions of the first 2 expert radiologists involved in this study as well as readers 1 and 2 were used. The ratings of reader 3 were left untouched to have a completely unseen data set for evaluation. Disagreements in classification between reader 1 and reader 2 were resolved by consensus in a joint reading session to construct the final training data set for the CAD algorithm. Examples of nodules for which the radiologists initially disagreed are depicted in Figure 5.3.

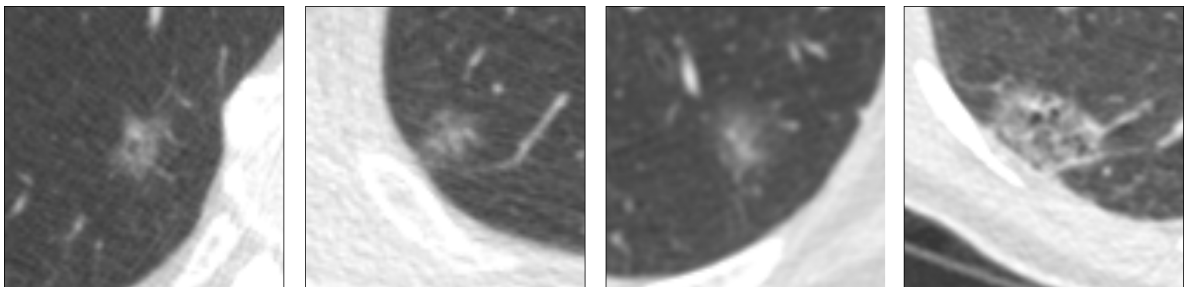


Figure 5.3: Examples of nodules for which a consensus meeting was needed and which were erroneously classified by CAD. Each subimage displays a transversal field of view of 6 x 6 cm. The left 2 nodules are nodules classified as part-solid by the radiologists but classified as non-solid by CAD. The right 2 nodules are nodules classified as non-solid by the radiologists but classified as part-solid by CAD.

Validation and Statistical Analysis

The performance of the CAD system to differentiate between solid, part-solid, and non-solid nodules was assessed using 2 analyses. First, we compared the rate of agreement between the human readers and CAD with the rate of agreement seen among the human readers using the Cohen κ statistics[33]. Confidence intervals were estimated using bootstrapping with 1000 iterations. Second, we assessed the performance of CAD to classify the lesions into solid, part-solid, and non-solid using the 2 independent readers, reader 3, and reader 4 as reference.

The CAD tool contains a statistical classifier used to predict the nodule type of a lesion given the features obtained from the segmentation-based lesion analysis. Sample lesions are needed to train the classifier to recognize features that describe the 3 different nodule types. Lesions that have been used in training and thus have been already exposed to the statistical classifier will subsequently be analyzed more precisely if analyzed again by the same classifier. Because the scores of readers 1 and 2 are used for training, a positively biased agreement between CAD and readers 1 and 2 would be found. To avoid an impact of this lesion-specific training on the CAD performance and decrease the risk for statistical overestimation, we used a leave-one-out cross-validation approach in our evaluation experiment, meaning that, for the classification of 1 particular lesion, the statistical classifier was retrained using the remaining nodules and thus did not consider any information with respect to the nodule under evaluation.

Agreement between human readers		Agreement between CAD and human readers	
Reader 1 vs Reader 2	0.70 (0.59 - 0.80)	CAD vs Reader 1	0.61 (0.50 - 0.72)
Reader 1 vs Reader 3	0.81 (0.71 - 0.90)	CAD vs Reader 2	0.72 (0.61 - 0.81)
Reader 1 vs Reader 4	0.63 (0.52 - 0.74)	CAD vs Reader 3	0.54 (0.42 - 0.66)
Reader 2 vs Reader 3	0.67 (0.56 - 0.76)	CAD vs Reader 4	0.60 (0.47 - 0.70)
Reader 2 vs Reader 4	0.75 (0.64 - 0.84)		
Reader 3 vs Reader 4	0.56 (0.45 - 0.66)		

Table 5.1: Agreement between human readers and between CAD and human readers for classifying nodules into solid, part-solid, or non-solid. Cohens κ statistics with 95% confidence intervals are reported.

5.3 Results

Pairwise agreement between each of the 3 human readers and CAD was moderate to good, with κ values between 0.54 and 0.72. Pairwise agreement within the 4 human readers alone was comparable, with κ values between 0.56 and 0.81. Table 5.1 lists all κ values.

Tables 5.2 and 5.3 show contingency tables for the performance of CAD versus the independent readers: reader 3 and reader 4. If the scores of reader 3 are used as reference, CAD correctly classified 98 of the 138 nodules (71%); 27 of the 40 disagreements (68%) referred to the differentiation between part-solid and non-solid, and 1 disagreement referred to the differentiation between solid and non-solid. The remaining disagreements referred to the differentiation between solid and part-solid. The nodule that CAD classified as non-solid, but for which reader 3 classified as solid, referred to a small 5-mm lesion in the lower right lobe (Fig. 5.4). Examples of other nodules where CAD and reader 3 disagreed are also depicted in Figure 5.4. If the scores of reader 4 are used as reference, CAD correctly classified 101 of the 138 nodules (73%); 24 of the 37 disagreements (65%) referred to the differentiation between part-solid and non-solid, and the remaining disagreements referred to the differentiation between solid and part-solid. CAD did not classify any solid nodule as non-solid nodule or vice versa.

These contingency tables show that the main difficulty is in differentiating part-solid from non-solid lesions. CAD classifies 16 nodules as part-solid, which are scored as non-solid by reader 4, but in return, CAD classifies 19 nodules as non-solid, which are scored as part-solid by reader 3. There is also disagreement in differentiating solid and part-solid lesions but to a lesser extent than that in differentiating part-solid and non-solid. These results show the difficulty of the nodule classification task at hand for both human experts and CAD, particularly differentiation

CAD / Reader 3	Non-solid	Part-solid	Solid	Total
Non-solid	16	8	0	24
Part-solid	19	46	6	71
Solid	1	6	36	43
Total	36	60	42	138

Table 5.2: Contingency table for reader 3 versus CAD.

CAD / Reader 4	Non-solid	Part-solid	Solid	Total
Non-solid	28	16	0	44
Part-solid	8	36	5	49
Solid	0	8	37	45
Total	36	60	42	138

Table 5.3: Contingency table for reader 4 versus CAD.

between part-solid and non-solid lesions.

5.4 Discussion

Nodule CT morphology has been advocated as an imaging biomarker for predicting the risk for harboring an invasive malignant tumor component[92–95]. The importance of nodule morphology for nodule management is also reflected by the current recommendations of the Fleischner Society for follow-up of intrapulmonary nodules[50]. Lesion size and lesion density as well as the presence and size of the solid component in part-solid lesions determine nodule management with respect to noninvasive follow-up or invasive diagnostic procedures. Visual human assessment of nodules is subject to interobserver variability even among experts[89,90], and this is also shown by the results of our observer experiment. In addition, if screening of large cohorts of high-risk subjects will be implemented, an increasing automation of the image reading and interpretation process may be the only option to keep costs reasonably low. These aspects provide the rationale to develop an automatic CAD tool for categorizing nodules into solid, part-solid, and non-solid, following the morphologic criteria suggested by the Fleischner Society. The presented CAD tool still requires a seed point; therefore, human input is still needed at this point. It has to be noted that this CAD tool closely follows established visual criteria for differentiating a solid from a non-solid part. To overcome the difficulty of drawing a line between solid

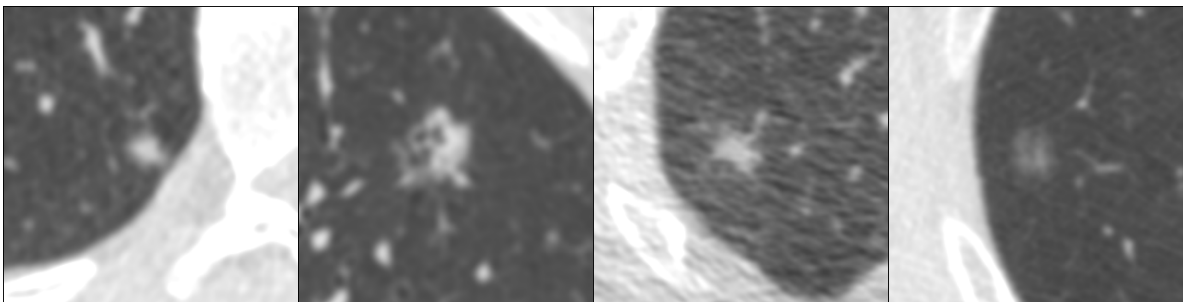


Figure 5.4: Examples of nodules that were differently classified by reader 3 and CAD. Each subimage displays a transversal field of view of 6 x 6 cm. Left, nodule classified as non-solid by CAD but as solid by reader 3. Middle left, nodule classified as solid by CAD but as part-solid by reader 3. Middle right, nodule classified as part-solid by CAD but as non-solid by reader 3. Right, nodule classified as non-solid by CAD but as part-solid by reader 3.

and non-solid areas purely on the basis of density thresholds, we incorporated additional features on the basis of local texture and histogram analysis, thus quantifying aspects of the lesion that may be qualitatively appreciable by visual analysis but are certainly not quantifiable by visual analysis alone.

In this study, we found a similar level of agreement between CAD and the human readers compared with the agreement among the human readers for classifying pulmonary nodules into solid, part-solid, and non-solid (κ ranges, 0.54-0.72 and 0.56-0.81). These results are encouraging with respect to using CAD to reduce interobserver variability among readers. Other studies involving CAD detection systems have already shown that CAD can reduce variability between radiologists[96]. Furthermore, these results could potentially pave the way for efficient analysis of large data sets. Three of the 4 readers were experienced chest radiologists, and the fourth one was involved in the screening trial and, as such, had gained special experience in the judgment of nodular lesions. Although we could not test whether CAD can exceed the performance of human readers because no ground truth on the nodule type is available, the performance of CAD was found to be comparable with expert opinions. Further studies are needed to compare the CAD performance with more inexperienced readers and the impact of a CAD tool on reader opinion when used as a secondary stand-by tool.

Our study suffers from a number of limitations. First, histopathological information was not available for all lesions. Because we developed our CAD tool on the basis of visual criteria used to guide followup management, we consider the lack of histopathology less relevant. However, for further development of CAD with potential to overrule human visual analysis, histopathological confirmation of training

data would be needed. Such studies are essential to solve remaining discrepancies between CAD analysis and visual expert analysis and to determine other quantifiable CT features that further increase the discriminating power between invasive and noninvasive tumor components. The work by Lee et al[97], which investigated the correlation between the presence of a solid core and an invasive component on pathology, and the work by Maldonado et al[98] and Kawata et al[99], which investigated the correlation between quantitative CT features and histopathological characteristics, show the potential of CAD analysis in this area. The work by Yanagawa et al[100] showed that solid volume and percentage of solid volume were significant indicators of lower disease-free survival. The CAD algorithm presented in this article also measures these indicators; therefore, we investigated whether these measures are significantly different between the 3 subgroups: solid, part-solid, and non-solid. Unpaired 2-tailed t-tests were performed, and a P-value below 0.05 was considered significant. In Table 5.4, the results are reported. The statistical tests showed that the solid percentage is significantly different between the 3 subgroups ($P < 0.001$, $P < 0.001$, and $P < 0.001$) and that the solid volume was significantly different between the non-solid and part-solid ($P < 0.001$) as well as between non-solid and solid ($P < 0.001$). Solid volume is significantly different between solid and non-solid as well as between part-solid and non-solid. Small solid volumes were found for the non-solid nodules where you would expect no solid component at all. This may be caused by noise or by the presence of vascular structures. Although the segmentation algorithm is able to exclude vasculature outside of the margin of the segmentation, it is not able to remove vessels within the margins of the nodule. Therefore, the algorithm is able to separate the solid core from vessels within the non-solid part of the lesion, but it may have difficulties in removing vessels from a pure non-solid lesion. Given the results of the statistical tests, the measurements of this CAD algorithm could potentially be used as a prognostic indicator of disease-free survival. This is, however, beyond the scope of this article.

Second, we did not assess intraobserver variability, which is, however, likely to be inferior to interobserver variability. Third, any analysis of histogram features using CT data is influenced by the technique of CT acquisition and reconstruction; it remains open how CAD would perform for CT data acquired differently and whether adaptations would be necessary to achieve comparable results. Fourth, the data in this study were obtained from a single screening trial and only 2 centers. Further research is needed to validate the system in data from other populations and acquired using different scanner protocols.

In conclusion, a novel automated classification tool for pulmonary nodules has been developed and validated on a large data set of screening detected lesions. The

Parameter / Subgroup	Solid volume [mm^3]	Solid percentage [%]
Non-solid	$29.4 \pm 41.5^{**\dagger\dagger}$	$4.28 \pm 6.41^{***\dagger\dagger\dagger}$
Part-solid	$669 \pm 1.66 \times 10^3$	$17.9 \pm 16.1^{***}$
Solid	$782 \pm 1.69 \times 10^3$	$49.7 \pm 16.2^{\dagger\dagger\dagger}$

Table 5.4: Quantitative measurements for the three subgroups. Data show mean $\pm$ standard deviation. Unpaired, two-tailed t-tests have been conducted to show statistical significance.
Data are expressed as mean $\pm$ standard deviation.
*** denotes statistical difference with the Solid subgroup at $P < 0.001$.
** denotes statistical difference with the Solid subgroup at $P < 0.01$.
††† denotes statistical difference with the Part-solid subgroup at $P < 0.001$.
†† denotes statistical difference with the Part-solid subgroup at $P < 0.01$.

performance of CAD was found to be equivalent with that of experienced chest radiologists. Further studies are warranted to assess the value of CAD as a supportive tool to select the appropriate workup for pulmonary nodules and the role of automatic nodule classification algorithms as part of risk models for predicting malignancy of pulmonary nodules.

Acknowledgments

The authors thank the investigators of the NELSON trial for providing data for this study.

Detection of interval change

6

Abstract

Detection of change between consecutive low-dose CT images is crucial in lung cancer screening. Visual comparison of CT scans is tedious and hence, automatic detection of change may aid human readers. In this study, we developed an automatic system for detecting change between low-dose CT images using analysis of subtraction images. Given two CT scans, a prior and a current scan, lung segmentation and non-rigid registration between the two scans is performed and a subtraction image is obtained by subtracting the deformed prior scan from the current scan. In the subtraction image, potential candidate regions with true change are determined using banded thresholding and connected component analysis. For each candidate, a set of intensity, shape and context features is computed. A GentleBoost classifier is used to classify candidate regions into true change or false positive regions. In total, 174 scan pairs were collected from a large lung cancer screening trial. An experienced radiologist annotated all relevant changes by inspecting the subtraction images and the two original images side-by-side. In addition, the quality of the subtraction images, an indication of the performance of the registration algorithm, was scored on a 1-5 scale. Performance was evaluated using free-response operating characteristic analysis. The quality of the subtraction images was rated high: only five subtraction images (4%) had a rating lower than 4. FROC analysis showed that the automatic system can detect 72% of all relevant change at an average of 2.0 false positives per scan. Thus, automatic detection of relevant change between low-dose CT images is feasible and may be of additional value when reading follow-up scans in a lung cancer screening setting.

6.1 Introduction

Lung cancer is the most deadly cancer in both men and women[1]. The National Lung Screening Trial (NLST), a randomized control trial in the U.S. including more than 50,000 high-risk subjects, showed that lung cancer screening using annual low-dose computed tomography (CT) reduces lung cancer mortality by 20% in comparison to annual screening with chest radiography[10]. Based on this result and subsequent modeling exercises to translate the NLST findings to population-wide screening scenarios, several organizations in the U.S. have started to endorse lung cancer screening. In 2013, the U.S. Preventive Services Task Force (USPSTF) gave low-dose CT screening a grade B recommendation for high-risk individuals and early 2015, the U.S. Centers for Medicare and Medicaid Services (CMS) has approved CT lung cancer screening for Medicare recipients. This means that lung cancer screening is reimbursed in full by private insurance companies and Medicare in the U.S. for eligible subjects. A person is eligible to enter a lung cancer screening program when fulfilling the necessary eligibility criteria such as being asymptomatic, between 55 and 77 years old, tobacco smoking history of at least 30 pack-years, and either current smoker or quit within the last 15 years.

Interpretation of screening CT scans is a labor-intensive task for radiologists. Early lung cancer manifests as pulmonary nodules which can be as small as 3 mm[12]. These small round structures have to be found in a large set of thin-slice CT images which typically have a section thickness between 0.5 and 2 mm. Next to the search for pulmonary nodules, the radiologist will also assess other abnormalities in the image and verify whether there are interval changes compared to the prior CT. These abnormalities and/or interval changes may be indicative for lung cancer, but can also be indicative for other diseases. Interval change between a new and a prior scan warrants close inspection by a radiologist. Computer-aided detection (CAD) systems have been shown to increase reader sensitivity and reducing reading time for detection of pulmonary nodules[19,24]. However, no automatic systems have been published which detect interval change on consecutive CT images. Since subjects in a screening setting will be scanned annually starting from the age of 55 years old, there will practically always be a prior scan available. In a screening scenario, a radiologist typically only checks for differences between a new and a prior scan if something suspicious is found on the new scan. Therefore, an automated analysis which exhaustively compares every region in the full lungs between new and prior scan could potentially aid the radiologist in the detection of relevant interval change.

In this study, we develop an automatic system for detection of interval change on consecutive low-dose CT scans. A fully automatic non-rigid registration method

is employed to register the current and prior scan. Using the deformation field which results from the registration method, the prior image is deformed to match the current image and subsequently, a subtraction image is obtained by subtracting the prior from the current image. Using the current, prior, and subtraction image, candidates potentially representing interval change are identified and a supervised classification scheme is employed to differentiate real interval change from false positive detections, such as registration artifacts, cardiac motion or breathing artifacts. Hence, we do not solely focus on detection of nodular changes, but on any interval change which may be present.

This paper is organized as follows. The data and the methods used for this study are explained in Section 6.2. In Section 6.3, the results are presented. Finally, in Section 6.4, we discuss the results, elaborate about the implications of our study and draw conclusions.

6.2 Methods

Data

Data for this study were collected from a large lung cancer screening trial. We selected all subjects who had a positive screening test and consequently were referred to the pulmonologist, and had at least one prior scan available. For these subjects, we collected the CT scans on the basis of which they were referred, and the scan before that. In these scan pairs, we expect an interval change present which caused the screening radiologists to refer them to a pulmonologist. In a lung cancer screening trial, subjects could be sent back into the screening when for example an abnormality turned out to be benign or if the subject refused the treatment proposed by the pulmonologist. Therefore, one subject can be referred to the pulmonologist multiple times and hence can have multiple scan pairs in our data set. In total, 124 scan pairs from 114 subjects were collected. The median number of days between the two scans was 363 days. This data set was evenly split into a training and test set while making sure that all CT scans of one subject were in the same set. As a result, 62 scan pairs were put in the training set and 62 scan pairs were put into the test set. In addition, 50 random pairs of CT scans for which long-term follow-up was scheduled were selected as additional normal cases for the test set on which no or minimal interval change is expected. In summary, the training set consisted of 62 scan pairs from 55 subjects, and the testing set consisted of 112 scan pairs from 109 subjects.

All screening CT scans were made at full inspiration with a 16-detector row CT scanner (MX8000 IDT or Brilliance 16; Philips Medical Systems, Cleveland, Ohio)

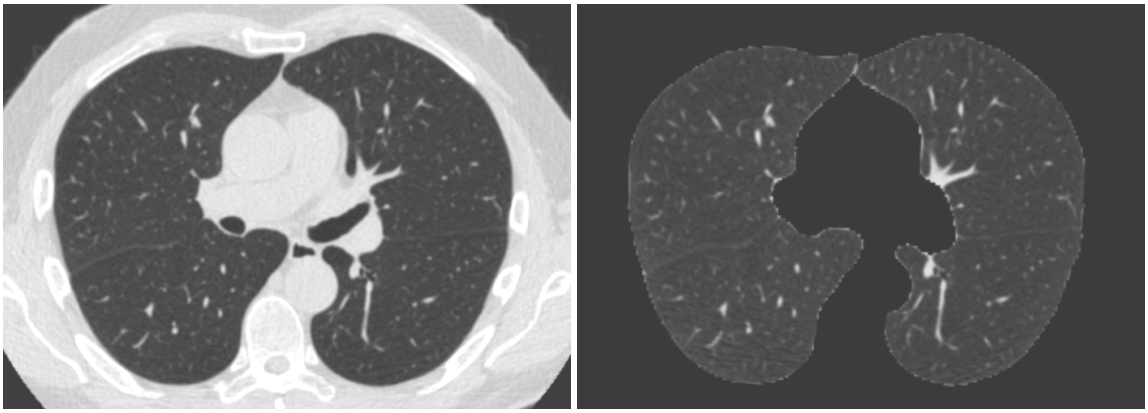
using a moderately soft reconstruction kernel (B; Philips Medical Systems), acquired in helical mode with 16 $\times$ 0.75 mm collimation. Exposure settings were 30 mAs at 120 kVp for patients weighing less than 80 kg and 30 mAs at 140 kVp for those weighing more than 80 kg. Axial images of 1.0-mm thickness were reconstructed at a 0.7-mm increment with a 512 $\times$ 512 matrix. In-plane voxel sizes varied from 0.53 mm to 0.88 mm.

Registration

A non-rigid registration algorithm dedicated for thoracic CT scans was used in this study[101]. This algorithm competed in the EMPIRE10 challenge, a lung CT registration challenge organized at the Grand Challenge workshop at the MICCAI conference in 2010, and ranked third out of 28 algorithms with an average runtime of only 110 seconds[102].

The method consists of a preprocessing phase and a non-linear lung registration phase. In the preprocessing phase, the lungs are segmented from both CT scans using a previously published segmentation algorithm[52]. Then, the lung regions are masked and the image is cropped around the lungs (see Fig. 6.1). This process ensures the removal of unwanted and potentially misleading image information from the registration process. Especially the removal of the rib cage is important. Due to the sliding motion of the lungs along the ribs during breathing, the ribs may be in a different position relative to the lung structures in the second scan which would result in high local deformations when registering two scans without masking out the lungs. The lung masks are subsequently smoothed, downsampled and their centers of gravity are aligned. The smoothed masks are registrered affine-linearly using the sum of squared differences (SSD) distance measure. The result is the initial deformation used for the next step; non-linear registration.

In the non-linear registration step, a transformation y is sought which maps corresponding structures in the two images onto another. This is modeled by an objective function $J(y)$, called the joint energy functional. Typically, the objective function of a registration process is built up from distance terms describing image (dis-)similarity and a regularizer which favors plausible solutions. Each term of the objective function has its own weight and the weights can be tuned to make certain terms more dominant than others. The lower the objective function $J(y)$, the better the two images match according to the model. Therefore, the registration process is an optimization problem where we minimize $J(y)$. The objective function $J(y)$ is formulated in Equation. 6.1.



(a) Current scan (b) Current scan, lung regions masked

Figure 6.1: Illustration of the masking of the lung regions. Only the lung regions are used in the registration process.

$$J(y) = D^{NGF}(R, T(y)) + \alpha S^{curv}(y) + \beta B(y) + \gamma V(y) \tag{6.1}$$

The first term, $D^{NGF}(R, T(y))$, with R being the fixed reference image and T the moving template image, is the main term of the objective function and provides most information to guide the registration. The lung regions in a CT scan consist of a rich structure of predominantly lung parenchyma, vessels and airways. The lung parenchyma covers most of the lung regions but exists of large homogeneous areas which provide little information to guide the registration process. Moreover, follow-up CT scans may have a slightly different inspiration level which will lead to attenuation differences in the lung parenchyma. The vessels in the lungs have a high contrast with the lung parenchyma and hence create strong edges in the image. Therefore, it seems sensible to focus on edges rather than intensities and hence, Normalized Gradient Fields (NGF) are used as a distance measure.

Since all structures except for the lungs are removed prior to the registration, the remaining deformation is assumed to be very smooth. Therefore, a curvature regularizer is chosen which penalizes second order derivatives of the deviation with respect to the initial deformation field after the affine-linear preregistration step. This is the second term, $S^{curv}(y)$ of the objective function, with the corresponding weight α.

The term $B(y)$ with weight β is to ensure that no singularities or extreme volume changes shall occur. The curvature regularizer does not safeguard against physically implausible deformations such as extreme volume expansion/shrinkage or foldings of the underlying grid. The term $B(y)$ penalizes deviations of the Jacobian from 1, meaning local volume expansion or shrinkage. The Jacobian is the determinant

of the Jacobian matrix, which is the matrix of all first-order partial derivatives of a vector-valued function, the deformation field in this case.

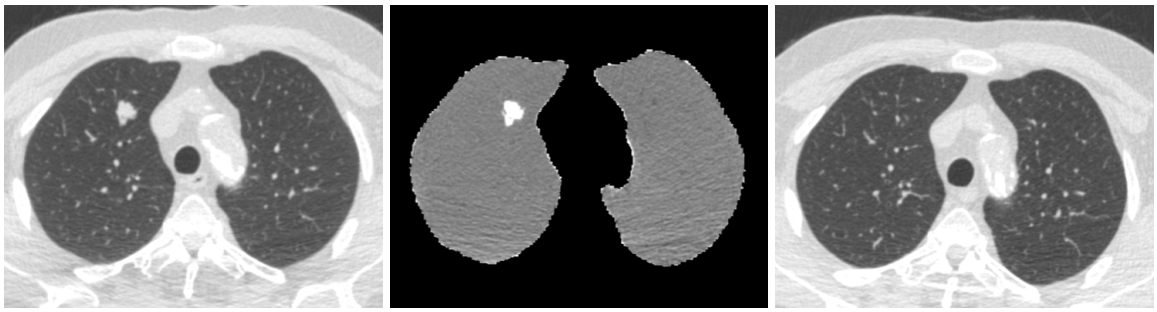
Finally, the last term $V(y)$ with weight γ is an additional term which incorporates lung boundary information in the model. It is reasonable to expect that the correct deformation should align the lung masks. The binary images of the segmentation masks of both scans are used and the sum of squared differences (SSD) is used as a penalty term.

Each of the weights of the terms were empirically determined and set to α = 5, β = 1, and γ = 1. The whole approach is embedded in a multi-level setting ranging from a coarse to fine resolution. Four levels are defined where the finest level has a deformation resolution of 64×64×64. Downsampling steps are applied for coarse resolutions which consist of Gaussian smoothing and a reduction of a factor 2 in all dimensions per step. For anisotropic data, the data are first downsampled in the dimension with the smaller voxel sizes to acquire isotropic resolution.

More details of the registration algorithm can be found in the publication by Rühaak et al.[101].

Subtraction images

Using the deformation which results from the registration method, the template image (the prior image) can be deformed to match the fixed image (the current image). Then, a subtraction image is obtained by subtracting the deformed prior scan from the current scan. In a subtraction image, new soft-tissue components will show up as white regions and disappearing soft-tissue components as black regions. An example subtraction image in which a new nodule has appeared is depicted in Fig. 6.2.

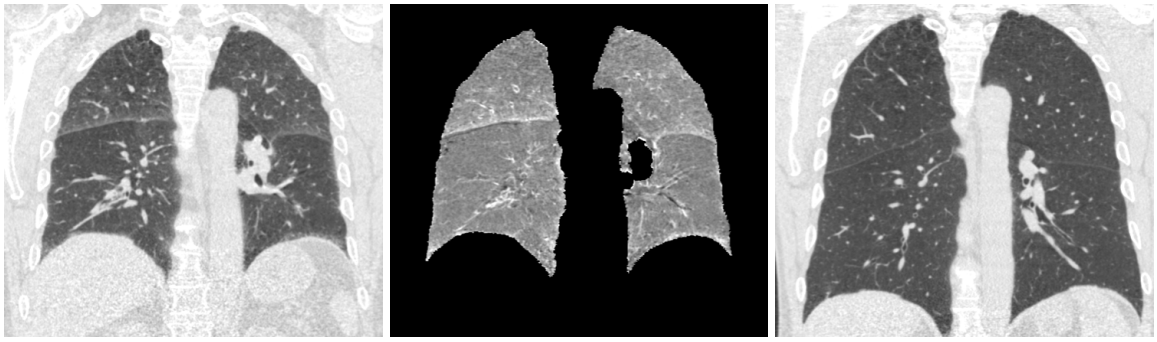


(a) Current scan (b) Subtraction image (c) Prior scan (not deformed)

Figure 6.2: Example of a result of the subtraction process. Images show transverse slices at the same magnification level. Original images are shown at window level setting -600/1600 HU. Subtraction image is shown at window level setting 0/1000 HU.

Automatic detection of interval change

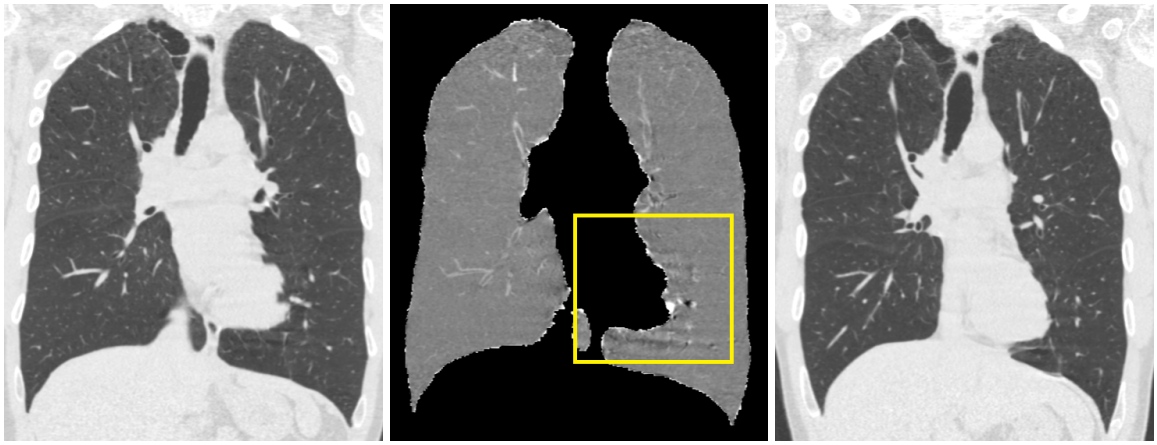
Based on the subtraction image and the two original images, an automatic system for detecting interval change was developed. In the subtraction image, potential candidate regions with true change are determined using banded intensity thresholding ($\leq$-250 or $\geq$250) and connected component analysis. This procedure finds black and white regions in the subtraction image separately. Small regions are removed for analysis, which are defined as regions with a volume smaller than 50 mm^3 (volume of an ideal sphere with a diameter of 5 mm). For each candidate, a set of intensity, shape and context features is computed. The intensity features consist of histogram statistics of the intensity values within the candidate region and in a $30 \times 30 \times 30$ mm cubic neighborhood around the candidate region, both computed on the subtraction image. Furthermore, histogram statistics of the intensity values of the full subtraction image are used. These features are added to capture information on the general quality of the subtraction image, but also to catch inhalation differences. If the inhalation level of the current scan is lower than the prior scan, less air is present in the lungs, leading to higher attenuation values within the lung parenchyma, generating positive values in the subtraction image. An example of two scans with a substantial inspiration level difference and the resulting subtraction image is depicted in Fig. 6.3.



(a) Current scan (b) Subtraction image (c) Prior scan (not deformed)

Figure 6.3: Example of a case where the new scan (left) has not been obtained at full inspiration, contrary to the scan protocol. Images show coronal slices at the same magnification level.

Since the shape of regions corresponding to interval changes between scans is not defined, only two basic shape features are added to the feature set: the volume of the candidate region in mm^3 and sphericity. In order to calculate the sphericity, a sphere S is defined at the center of mass of the candidate region with the same volume as the candidate region. Then, sphericity is defined as the ratio between the volume of the voxels of the candidate region within sphere S and the total volume



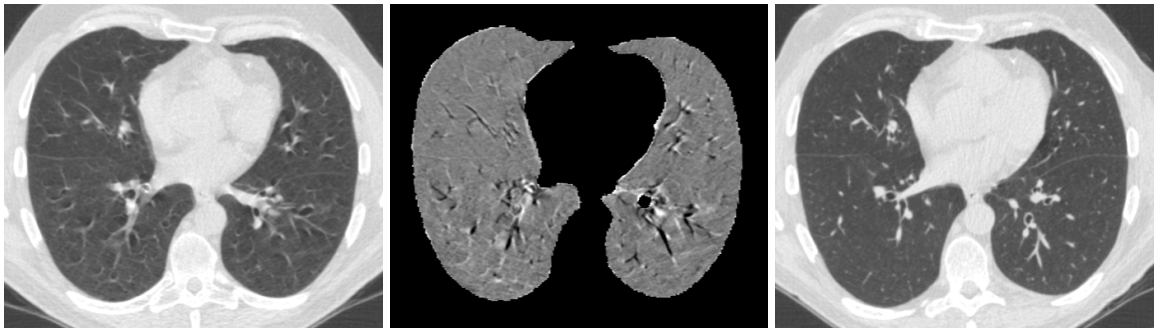
(a) Current scan (b) Subtraction image (c) Prior scan (not deformed)

Figure 6.4: Example of artifacts in the subtraction image caused by cardiac motion. Cardiac pulsation caused a staircase artifact that is too extreme to be corrected by the deformation field. Images show coronal slices at the same magnification level.

of sphere S. Finally, a set of context features are added. Context features describe the location of the candidate region with regard to other structures and other candidate regions and have shown to be of great value in other computer vision tasks[72,77]. Misregistration of structures typically leads to a black-white pattern in the subtraction image. Therefore, the number of black and white regions in the neighborhood of a candidate region provides information about the local level of misregistration. This is also captured by the intensity features but is encoded more explicitly in the context features. Furthermore, the location of the candidate region within the lungs is encoded. Since cardiac gating is not used for the CT scans used in this study, pulsation artifacts are present in the original CT scans which leads to artifacts around the heart in the subtraction image (see Fig. 6.4).

In summary, the set of context features consists of (1) the number of candidates in a $30 \times 30 \times 30$ mm neighborhood, (2) the number of candidates in a 10 mm transverse section, (3) the distance of the candidate region to the lung boundary, (4) distance to the center of gravity of the lung segmentation, and (5) the relative X,Y,Z location of the candidate region. The second context feature, the number of features in a 10 mm transverse section, is added to prevent false positives caused by breathing artifacts. If a patient shortly breathes during the CT examination, a small transverse section of the CT scan will show breathing artifacts which are not present in the other CT scan. This leads to many artifacts in the subtraction image (see Fig. 6.5).

Based on this feature set, a GentleBoost classifier using 250 regression stumps is trained to differentiate real change from subtraction artifacts. Performance of the system was evaluated on the independent test set using free-response operating char-



(a) Current scan (b) Subtraction image (c) Prior scan (not deformed)

Figure 6.5: Example of artifacts in the subtraction image due to breathing motion. Breathing motion is visible in the new scan (left image). Images show transverse slices at the same magnification level.

acteristic (FROC) analysis.

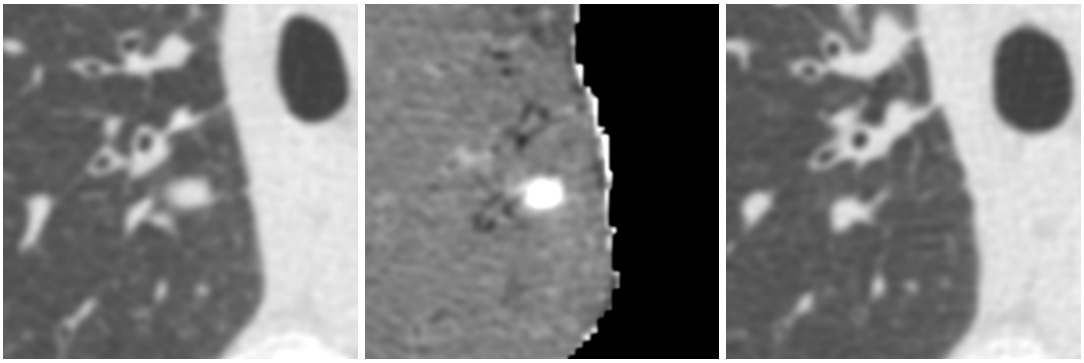
Annotation of relevant change

All relevant changes in all scan pairs were annotated by inspecting the subtraction images and the two original images side-by-side. A researcher with >4 years of experience in reading chest CT scans for nodule detection annotated the training set. An experienced radiologist annotated the test set. The readers were instructed to annotate all significant change. This excludes for example small nodules, small consolidations, minor collapse, and slight thickening of the fissure. In addition, the quality of the subtraction images, an indication of the performance of the registration algorithm, was scored by the radiologist on a 1-5 scale: (1) complete failure; no result, (2) severe global misregistration, (3) major local registration artifacts, (4) minor local registration artifacts, and (5) no registration artifacts visible.

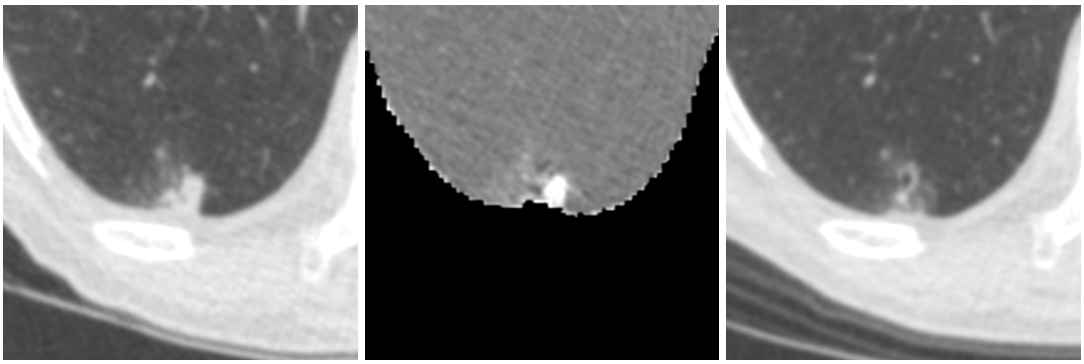
6.3 Results

In total, 94 relevant changes were annotated by the experienced radiologist in the test set. On the training set, 165 regions were annotated. The majority of the annotated changes consisted of appearing/disappearing or shrinking/growing nodules, appearing/disappearing mucus in airways, or onset of other opacifications. Fig. 6.6 shows examples of the annotated changes on the test set.

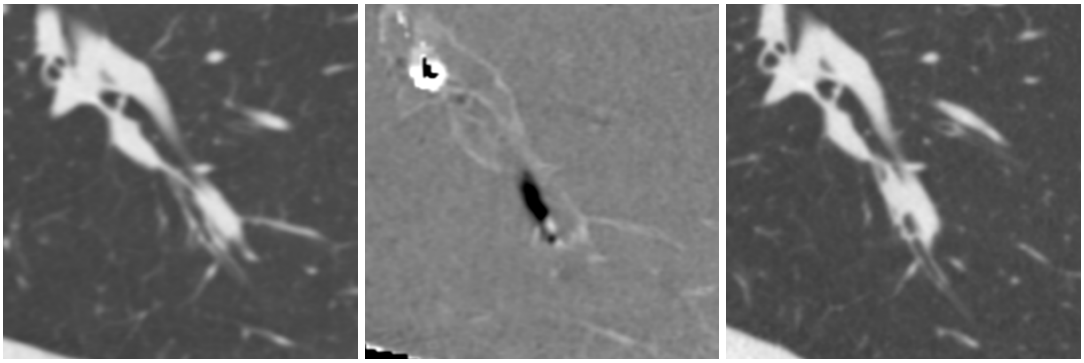
The quality of the subtraction images in the test set was rated high: only five subtraction images (4%) had a rating lower than 4. Three of the five cases with a lower score were caused by a failure of the lung segmentation on one of the CT



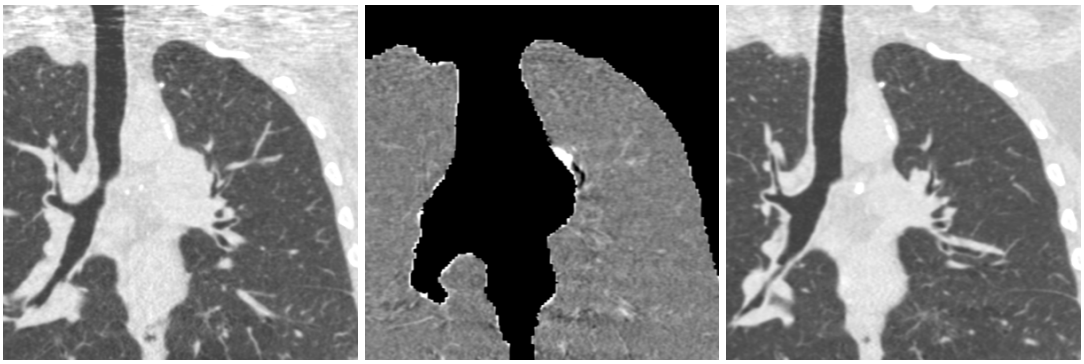
(a) New nodule



(b) Growing nodule

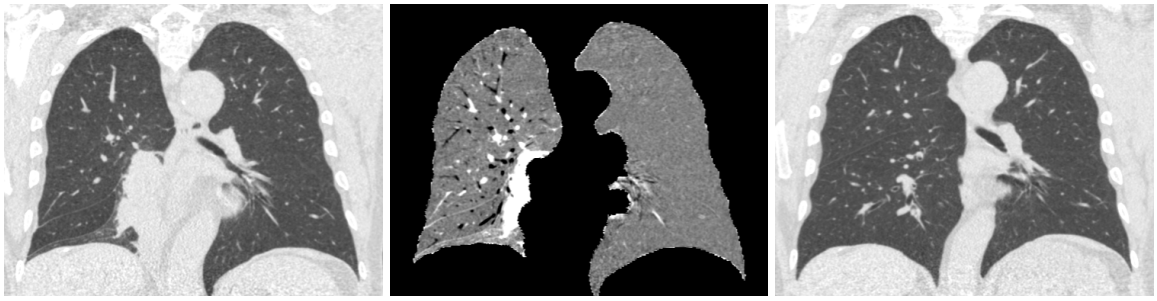


(c) Disappearing mucus



(d) Hilar enlargement

Figure 6.6: Examples of interval changes found by experienced radiologist. Each example consists of three subimages which show the new scan (left), the subtraction image (middle), and the prior scan (right).



(a) Current scan (b) Subtraction image (c) Prior scan (not deformed)

Figure 6.7: Example of major misregistration in right lung due to the rapid development of a large tumor. The registration model assumes a smooth deformation and therefore, the method fails to register correctly. Also note that part of the tumor is excluded from the lung segmentation. Images show coronal slices at the same magnification level.

scans. The other two lower scores were cases in which a large tumor developed, which caused major registration artifacts in the affected lung because the model was not able to model this large local deformation (see Fig. 6.7). FROC analysis showed that the automatic system can detect 72% of all relevant change at an average of 2.0 false positives per scan. The FROC curve is depicted in Fig. 6.8. Confidence intervals (95%) were estimated using bootstrapping with 5000 iterations.

Fig. 6.9 shows the three most suspicious false positives findings and three false negatives of the system. All three false positive findings are small new nodules which were not annotated by the radiologist. The three false negatives show a small new nodule missed in the candidate detection stage, a small growing solid nodule, and an appearing ground glass lesion. Both the small growing lesion and the appearing ground glass lesion received a very low probability by the classifier.

6.4 Discussion

Detection of interval change between consecutive low-dose CT images is crucial in lung cancer screening. Manual comparison of CT scans is tedious and hence, automatic detection of interval change may aid human readers. In this study, we have shown that automatic detection of relevant change is feasible. The system reaches a sensitivity of 72% at an average of 2.0 false positives per scan. The locations of the detected interval changes of the system can be marked on the CT scan as additional CAD prompts and hence attract the focus of the human reader to closely inspect that region on both current and prior scan. Future observer studies are needed to inves-

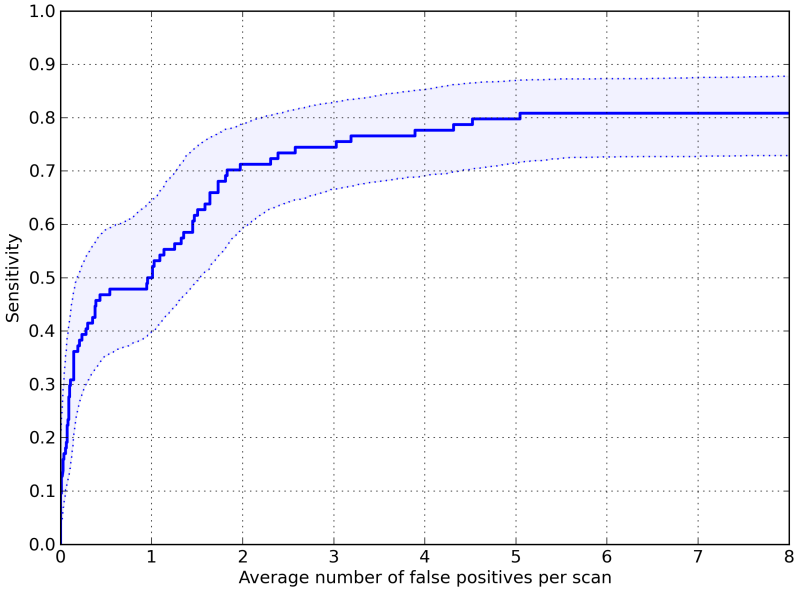


Figure 6.8: FROC curve for the proposed automatic detection system for interval change. Shaded areas around the curve indicate 95% confidence intervals, estimated using bootstrapping with 5000 iterations.

tigate whether this system can increase reader sensitivity for interval change detection. The analysis of the false positives and false negatives on the test set showed that the system finds additional (small) lesions which were not annotated by the radiologist. In addition, we showed two examples of interval changes which were missed by the system; minor growth of a small nodule and an appearing ground glass lesion. Underrepresentation of these types of interval change in the training data may be the reason why these interval changes received such a low probability. Hence, extending the training data set may solve this problem.

The system employs a non-rigid registration algorithm and based on this algorithm, subtraction images are created and analyzed. The quality of subtraction images was scored 4 or higher for the vast majority of the cases, implying that the subtraction images can be interpreted well by human readers. A previous study by Aoki et al.[103] already reported an increased performance for detection of new nodules when subtraction images are presented to human readers without considerably extending the reading time. Presenting subtraction images to human readers in addition to the normal CT scans may be another way to use the presented work, but adds to the number of images a radiologist has to inspect.

In this study, we did not investigate whether subtraction images are also suitable for accurately assessing nodule growth. A study by Staring et al.[104] showed that subtraction images improve the evaluation of subtle changes in pulmonary ground-

glass nodules and decrease reader interobserver variability. In that study, a local rigidity penalty term was employed to prevent the nodule from deforming in a non-rigid fashion because this could effectively conceal nodule growth. The registration method employed in our study also has a penalty term which prevents physically implausible deformations but we did not investigate to what extent the registration procedure would locally expand a growing nodule and thus create a subtraction image in which the growth is not visible. To detect growth or shrinkage in such cases, the Jacobian of the deformation field could be used as an additional feature, and the Jacobian could also be used in the candidate detection procedure. We expect however, that this is not needed, because the regularization term $B(y)$ largely eliminates local expansion or contraction. This is beyond the scope of this study.

The preprocessing step of the proposed algorithm includes a lung segmentation step which excludes all image information outside of the lungs. Since this information is excluded, we can assume that the remaining deformation is very smooth and this is an appealing property of the employed registration process. However, the hilar area is also excluded which is an important limitation of the current method. Since the deformation in the hilar area is also assumed to be relatively smooth, future experiments should include this area in the registration process. Then, we can evaluate whether interval changes in this area can also be detected using subtraction images. For example, enlargement of the hilum is an important finding which may indicate lymph node metastases and this may be detected at a higher sensitivity when using subtraction images. However, in the current proposed method, a large part of the hilum is typically excluded due to the lung segmentation. Another application where subtraction image may be of additional value is the interpretation of growth or shrinkage of metastases in the lung during treatment, or the evaluation of progression of interstitial lung disease.

In this study, we used CT scans from subjects who were referred to the pulmonologist. Consequently, most changes were relatively obvious and an extended validation with more subtle changes is necessary to investigate whether the proposed algorithm can also detect very subtle changes. The false positives presented in Fig. 6.9 show some missed small nodules which may suggest that this algorithm is also suitable to detect new small nodules.

In conclusion, we showed that automatic detection of relevant change between low-dose CT images is feasible and may be of additional value when reading follow-up scans in a lung cancer screening setting.

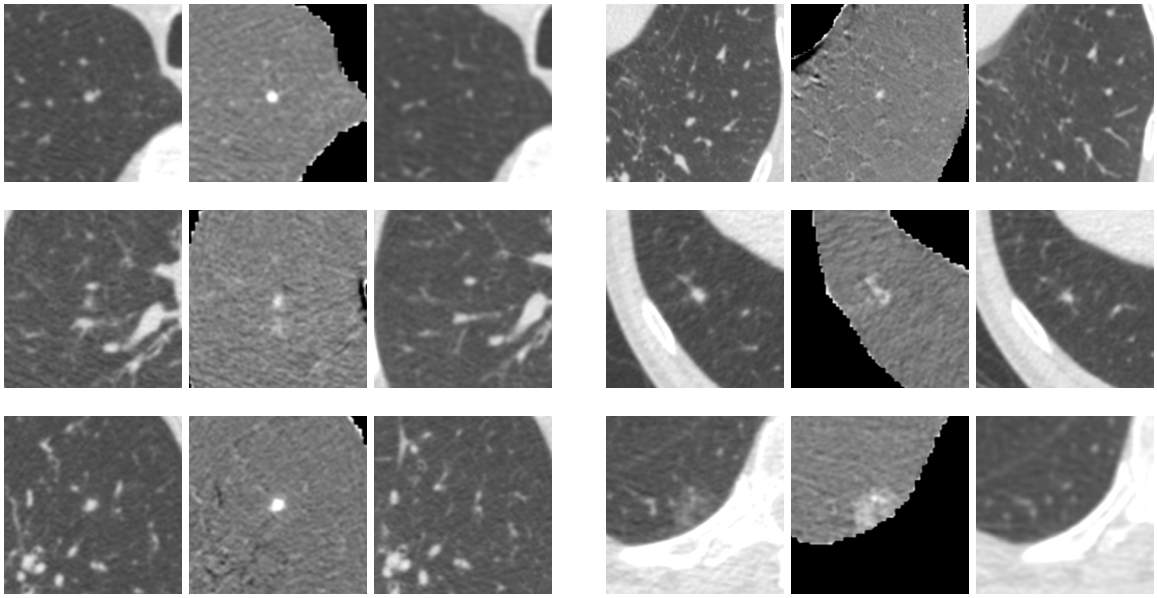


Figure 6.9: Examples of false positives and false negatives. Each example consists of three subimages which show the new scan (left), the subtraction image (middle), and the prior scan (right). The left column shows the three false positives with the highest likelihood. All three false positives are indeed true change, but the annotation protocol did not require the radiologist to annotate small new nodules. The right column shows three examples of false negatives of the system. The top example shows a small new nodule which was missed in the candidate detection. The middle example shows very subtle growth of a tiny solid nodule. The bottom example shows an appearing ground glass lesion. Both the middle and bottom example did have a candidate region but these received a very low probability by the classifier.

Acknowledgments

This work was supported by a research grant from MeVis Medical Solutions AG, Bremen, Germany and by a research grant from the Netherlands Organisation for Scientific Research, project number 639.023.207.

Integration into a screening workstation

7

Abstract

This thesis described several methods that could be of value if they were integrated into software that is used in clinical routine or in CT lung screening programs. Such software needs to be certified for use as a medical product. Integration of research algorithms into products is usually outside the scope of PhD research projects in medical image analysis, but the project described in this thesis is an exception to that rule. In this chapter I briefly describe the optimized reading workstation for chest CT scans that I developed in the last few years of my PhD project, and explain how the methods described in the other chapters of this thesis are integrated in this solution, that is currently commercially available as a medical product.

7.1 Translating research to the clinic

Last year, one of the leading Dutch newspapers, De Volkskrant, called the Dutch universities "PhD factories" (De Volkskrant, November 25, 2014). The number of dissertations "produced" by Dutch universities increased from less than 2,500 in 2000 to almost 4,500 in 2013. PhD students are responsible for most of the Dutch publication output. Universities receive money from the government for every doctoral degree awarded (90 kEuro per thesis produced) and this has resulted, according to many critics of the current system, in a preference for quantity over quality. In addition, the question is raised what the value for society is of this endless pile of PhD theses, such as the one you are reading right now.

Some Dutch universities have recently added a formal requirement to discuss how the research described in a thesis can be of benefit for society at large. An example is the University of Maastricht where an addendum regarding valorization must be part of each dissertation. Questions that may guide the PhD candidate in writing this addendum include[105]:

- (Relevance) What is the social (and/or economic) relevance of your research results?
- (Target groups) To whom, in addition to the academic community, are your research results of interest and why?
- (Activities/Products) Into which concrete products, services, processes, activities or commercial activities will your results be translated and shaped?
- (Innovation) To what degree can your results be called innovative in respect to the existing range of products, services, processes, activities and commercial activities?
- (Schedule & Implementation) How will this/these plan(s) for valorization be shaped? What is the schedule, are there risks involved, what market opportunities are there and what are the costs involved?

This chapter can be considered a personal attempt to answer some of these questions.

The main focus of this thesis was on the development of computer algorithms to extract relevant information from thoracic CT scans. Therefore, embedding these algorithms in actual software products, that can be used to analyze CT scans in either lung CT screening programs, or clinical routine, is an obvious way to increase the

societal value of the research that led to this thesis. Doing this is in line with the mission of the Radboud University Medical Center which is *to have a significant impact on healthcare*. It was also of interest to MeVis Medical Solutions AG (Bremen, Germany), the company that provided funding for my research. During my work in the last five years, I have actively worked on the translation of my research algorithms into a medical product. I did this next to my actual research work and this is in fact one of the reasons my PhD trajectory took a bit longer than the normally allocated four years. In this chapter I will briefly explain some characteristics of this process of translating research from the lab into a clinical product, and I will describe the product that is the result of this work. Interestingly, this description also serves as a summary of much of the contents of the other chapters in this thesis, and the reader can see how these separate pieces can be fitted together into a coherent solution for analyzing chest CT scans in a screening and/or clinical setting.

Why is research software usually not directly integrated in real products? Part of the answer is that translating research algorithms into products that can be used by clinicians is not a trivial task. This is often a surprise to clinicians who see research software in action or read a paper about it and do not understand why they can not use it immediately. I briefly mention several issues regarding integration, speed, robustness, and regulatory requirements.

Integration. The requirements for research software are completely different than for commercial software. Take for example the system for subsolid nodule detection described in Chapter 2. This consists of a sequence of steps (segmentation, candidate detection, feature extraction, candidate classification). It is not unusual that the researcher writes separate, independent pieces of software for the different steps, even using separate software systems. For a publication, it is only needed that data used in the experiments of the paper are processed, and this can be done in a stepwise manner. In a product, obviously a single piece of software is needed. Furthermore, integration of a piece of software into the clinical workflow of radiologists can be done in many different ways. For example, the result of an algorithm can be stored in a DICOM object and sent to PACS. Then, a radiologist can inspect the result in his/her usual workstation. However, since many different DICOM objects exist, one has to check the DICOM conformance statement of the local PACS system to be sure that the produced DICOM object is supported by that PACS system, and the same holds for the workstation. If there is user interaction involved, the software needs to be integrated into a workstation. If a dedicated workstation is developed by the research group, it needs to have all viewing capabilities which a radiologist typically uses, such as window level adjustment, zooming, etc. and one has to keep in mind that switching workstations is very inconvenient for a radiologist.

Speed. The runtime of an algorithm is not very important for a publication, and researchers often use multiple computers or special hardware to run their algorithms. But the user of a product will not be enthusiastic if it takes hours to analyze one scan, or if a large computation cluster is required to run the algorithm in an acceptable amount of time.

Data selection and robustness. When data is selected for a publication, it is common to use a wide set of exclusion criteria, and at this stage the researcher discards any data that the software system is likely to have problems with. Products should work with any input data that meets certain minimal quality requirements. In scientific studies, data often comes from a single center or from a few centers, but products should work with data from any center. Anyone intimately familiar with the DICOM standard will be able to testify that this is not as easy as it might seem. Research software may even often crash, and in that case the developer fixes an error, or modifies the input data, or simply removes the scans on which the software crashes from the study.

Professional software development. Open-source libraries are often used in research software, but some libraries cannot be used for commercial use. For example, if you use open-source libraries licensed under the GNU General Public License (GPL), you are required to pass on the same freedoms as you received, which means that you also have to make your software open-source. Evidently, this is usually not an option for commercial software. This needs to be considered while developing software which may end up in a commercial product later on. Furthermore, software should not crash in products and extensive tests are written to verify automatically, in a build system, that the software, while undergoing changes during development, keeps passing the tests. In our group, we have implemented a quality system for software development that meets the relevant regulatory requirements (MDD Annex II and ISO 13485). This was important in supporting the effort to make our research software gradually more integrated, better documented, fast and robust.

7.2 CIRRUS Lung Screening

In this section we describe the general workflow of our high throughput lung CT screening solution: CIRRUS Lung Screening. This software is designed to let a radiologist quickly and accurately report chest CT scans. The output is a standardized report. While reading a case, the user has direct access to all prior scans of the same subject, something that is not available in other products currently on the market. The software includes computer-aided detection and volumetric measurements for both solid and subsolid nodules, as recently recommended in a joint white paper by

the European Society of Radiology and European Respiratory Society on lung cancer screening[11]. The workstation is developed for a dual monitor setup. In the figures at the end of the chapter, some are from the left screen and some from the right screen. First we can distinguish *processing*, which takes place before the scan is presented to the reader, and the *reading workflow*. For both steps, the situation is simpler in case the scan to be read is the baseline scan (the first scan of a new subject). Additional steps are necessary if prior scans are available (which are always assumed to have been reported already) and these are indicated with an asterisk.

The *processing* consists of:

1. Verifying whether the input DICOM data is valid, whether the scan is a chest CT scan, and a quality check. This is accomplished by inspection of the DICOM headers. The quality check checks whether the input data is thin-slice (slice thickness ≤ 3 mm) and has at least 50 slices.

2. Verifying whether there are prior CT scans of the same subject. If yes, retrieve prior CT scans and the reports and reading results from the database.

3. Lung segmentation. This is a prerequisite for many algorithms. Two in-house developed algorithms are computed[52,106]. The second algorithm is used as a fallback in case the first algorithm fails.

4. Airway and vessel segmentation. This is needed for feature computation in the subsolid CAD system. Two in-house developed algorithms are used[53,54].

5. Lobe and segment segmentation. This is used to on-the-fly assign lobe or segment labels to nodules annotated by the human reader. Two in-house developed algorithms are used[107,108].

6. Running various CAD systems to detect nodules. The subsolid CAD system described in Chapter 2, and the ISICAD and Herakles systems from Chapter 4 are used in this step. The various results of the different CAD systems are merged into a list of possible nodules.

7. * Registration. The dedicated lung CT registration algorithm which was described in Chapter 6 is used to register a new scan with prior scans[101]. The output of the registration procedure, the deformation fields, are stored on disk such that they can be easily loaded by the workstation.

8. * Propagation of prior findings. All nodules which were marked on the prior scan are automatically propagated to the new scan using the registration algorithm. In this way, the evolution of nodules can be easily tracked.

9. * Merging the prior findings and CAD marks. CAD marks and propagated nodules may point to the same nodules and therefore, a merging step is applied in which CAD marks which are at the same location as propagated nodules are removed.

The *reading workflow* consists of:

1. Opening a case.

   The patient browser of the workstation lists all available studies and displays their state (see Fig. 7.1). Studies which are ready to be read have the state "Preprocessed". Studies which have been read and approved have the state "Approved" and are colored green. Upon opening a case, the CT scan of the new study is displayed on the left screen, and all prior scans can be viewed on the second screen. All scans and processing results are automatically preloaded to computer memory to ensure rapid opening of cases. When a complete subject has been loaded, the software already loads all files from the next subject in the worklist into memory, so that the user can quickly proceed to the next case when the current subject has been reported on. Each CT scan is shown using a 1-by-3 layout: a large view with three orthogonal views is used, see Fig. 7.2. The three orthogonal views contain orientation indicators which indicate the location of the displayed section in the main view. Each prior scan is accessible via the corresponding tab on the second screen, see Fig. 7.3. During processing, the new CT scan has been elastically registered to all prior scans and these results are used to enable linked scrolling between current and prior scans. The lung segmentation is used to automatically initialize the viewer at the top of the lungs.

2. Reading a case.

   The reader can start scrolling from the top to inspect the lungs for potential nodules. A list with all annotated findings is displayed in the left panel, see Fig. 7.8. In this list, propagated nodules and CAD marks are also displayed. The list gives a quick overview of all annotated findings and can be used to quickly navigate towards the findings - a click on the finding will center the viewers on the corresponding location. Each annotated finding can be described using the following characteristics (which are accessible through combo boxes below the nodule list): Type, Segment (or Lobe), Morphology, and Result. The possible values for these characteristics are based on the scoring forms used in the Pan-Canadian Early Detection of Lung Cancer Study (PanCan study)[109]. For Type, Solid, Semi-Solid, Ground-Glass, Calcified, and Perifissural can be indicated. Using the nodule classification explained in Chapter

5, Type is automatically set by the software when a new nodule is annotated, but can be adjusted by the reader using the combo box, if necessary. The segment in which the finding is located is also automatically filled in using the results of the lobe and segment segmentation. For Morphology, Well defined, Lobulated, and Spiculated can be indicated. This needs to be added manually by the user at the moment. In the Result combo box, Unchanged, Resolved, Benign, Smaller, Growing, Density Increase, or Resected can be selected. The following sections explain some aspects of the reading process in more detail.

CAD marks. For each case, a set of CAD marks is generated using the output of the CAD systems. When CAD is enabled, the CAD marks are added to the nodule list in the left panel. The workstation can be configured to automatically enable CAD for each case, or whether a user first has to press an "Enable CAD" button. CAD marks are automatically sorted by size, such that the most relevant CAD marks are listed first. If a false positive CAD mark is generated, the user can delete the CAD mark by pressing the "Delete" button. Annotated findings and CAD marks are displayed in the viewers using squares. Green squares are accepted findings: accepted CAD marks, or manually added findings. Blue squares are CAD marks which still need to be handled by the user. The ID of the findings is listed next to the squares. CAD marks have an ID which start with the letter C.

Segmentation. If a user finds a new nodule, a double-click in the viewer will generate a new finding at that location. A double-click on a CAD mark accepts the CAD mark as a nodule finding. Upon a double-click, the nodule is automatically segmented using an automatic nodule segmentation algorithm, which can handle both solid and subsolid nodules[57]. A segmentation contour is displayed to indicate the segmentation boundaries. In addition, the Analysis tab is activated on the second screen, see Fig. 7.7. The Analysis tab is designed to give the user a clear overview of the evolution of the lesion over time. It shows a grid of zoomed views of the nodule at a maximum of 5 timepoints. Each column represent one timepoint, and provides an axial, coronal, and sagittal zoomed view of the nodule. Above the orthogonal views, detailed information about volume, mass, density, and growth is displayed. For growth, the volume doubling time (VDT) and mass doubling time (MDT) are calculated. VDT is used in NELSON to discriminate positive, indeterminate, and negative screening cases[9]. Increase in mass has been reported to be an early indicator of growth for subsolid nodules[110]. If needed, manual correction of the segmentation is possible by adjusting the parameters of the automatic segmentation algorithm. If no satisfactory segmentation can be obtained by adjustment of

the segmentation parameters, a manual diameter can be drawn to measure the nodule size. Furthermore, for part-solid nodules, an extra segmentation of the solid core is performed, see Fig. 7.8 and Fig. 7.9. Solid core measurements are important for the management of nodules and are included in the recently released LungRADS guidelines[111].

Assessment of likelihood of malignancy. A predictive model for the likelihood of malignancy of nodules is included into the workstation and is automatically calculated for each nodule[88].

Retro nodules. If a user finds a new nodule, but notices that the nodule was already there at a prior scan, the software supplies a way to record this in the database. This is accomplished by supporting the annotation of nodules on prior scans, which we refer to as so-called retro nodules. The concept of retro nodules was introduced by the investigators of the PanCan study and facilitates a structured way to handle nodules already visible on prior CT scans[109]. A simple click on the Retro button in the analysis panel allows a user to annotate and segment a retro nodule.

Additional characteristics. In addition to annotation of nodular findings, the workstation also supports the annotation of additional characteristics which may be important in a lung cancer screening scenario. These additional characteristics are currently focused on the two other conditions for which heavy smokers are at risk: chronic obstructive pulmonary disease (COPD) and cardiovascular disease (CVD). In the Additional Characteristic panel, which is located below the Nodule panel, the following characteristics can be annotated (see Fig. 7.2): Emphysema (Extent, Type, and Distribution), presence of airway wall thickening, coronary artery calcification (separated into left main and left anterior descending artery (LMLAD), circumflex artery (CIR), and right coronary artery (RCA)), and lymph node involvement.

3. Reporting.

   After a user has finished reading, he/she navigates to the Report tab and a structured report is automatically generated by the software. On the second screen, the prior reports are also provided. The report provides general patient information, gives a quick overview of the annotated findings and their characteristics, and optionally provides follow-up recommendations according to Fleischner guidelines and/or LungRADS[111]. Based on the report, a user can decide upon the appropriate follow-up for the patient. The suggested follow-up and other case comments can be added to the report. Reports can be exported to PDF for distribution outside of the software. Finally, the user signs off the

case and the next case, which is already preloaded to computer memory in the background, is presented to the user to facilitate quick reading of a worklist.

7.3 Different versions of the software

Different versions of our software are available. The research version, entitled CIRRUS Lung Screening, runs the processing pipeline on our in-house cluster, is flexible, has lots of additional options for e.g. reader studies, and contains experimental features. The commercial version, Veolity (`http://www.veolity.com`), uses a client-server architecture, has communication with PACS systems integrated, can use PACS as storage database, divides worklists over multiple clients, and is fully certified.

Both versions are in active use at several sites in North America, Europe and Asia, and form the basis for research collaborations that build a network of clinical and technical partners that is needed to address the challenge of cost-effective and efficient lung CT screening.

7.4 Preliminary evaluation

We performed a preliminary evaluation of the feasibility of rapid reading using the described workstation[112]. In this study, we investigated the performance of rapid reading of chest CT scans with integrated CAD support, with the goal of quickly assigning a subject to either regular one-year follow-up, short-term follow-up or immediate work-up. From the baseline round of a large lung cancer screening trial, we randomly selected 23 cases from each of the three categories used in the trial: 1) no significant nodules, 1 year follow-up CT; 2) nodule 50-500 mm^3, 3 month follow-up CT; 3) nodule $>$500 mm^3, referral to pulmonologist. Seven blinded readers read all cases in random order in a single session as follows. First, CAD marks were inspected and accepted or rejected. Next, readers quickly inspected the scan and added relevant nodules if CAD had not identified these. Finally, readers assigned the scan to one of the three categories of the screening protocol. We showed that 73$\pm$7% of cases (range 58-80%) were assigned to the correct category. Total median reading time per case was 67$\pm$17 seconds. We concluded that with the support of highly effective CAD systems, nodule volumetry, and an optimized reading environment, it is possible to accurately read lung cancer CT scans in around one minute per case.

7.5 Outlook

The current version of our software is in active use at different sites and this leads to valuable user feedback which will help to improve our product. In the coming years, lung cancer screening will be implemented on a larger scale in the United States and this may lead to updated guidelines, new classification schemes, etc. Therefore, we have to make sure that our solution is flexible and can adjust to the needs of users. Future development will include a more sophisticated reporting engine, such that the layout of reports can be easily configured. Furthermore, automatic methods for the assessment and quantification of chronic obstructive pulmonary disease (COPD) and cardiovascular disease (CVD) will be integrated, such as quantification of emphysema, airway wall thickening, and coronary artery calcification[113]. The detection of micronodules presented in Chapter 3, and other lung CT analysis tools could be integrated as well. Other new features which are on our list are the integration of subtraction images presented in Chapter 6, better automatic classification of nodule type, and inclusion of patient history into our database.

7.6 Conclusion

Developing a product requires different skills compared to doing research, but is a rewarding effort. A product handles a complete workflow and designing that workflow, choosing which features to add and which to omit, is a challenging task. However, it is crucial to facilitate the use of the methods described in this thesis in clinical practice. The solution for high throughput lung CT screening described in this chapter is a strong basis for future research into the role of automated analysis in lung cancer screening.

IMAGE GALLERY

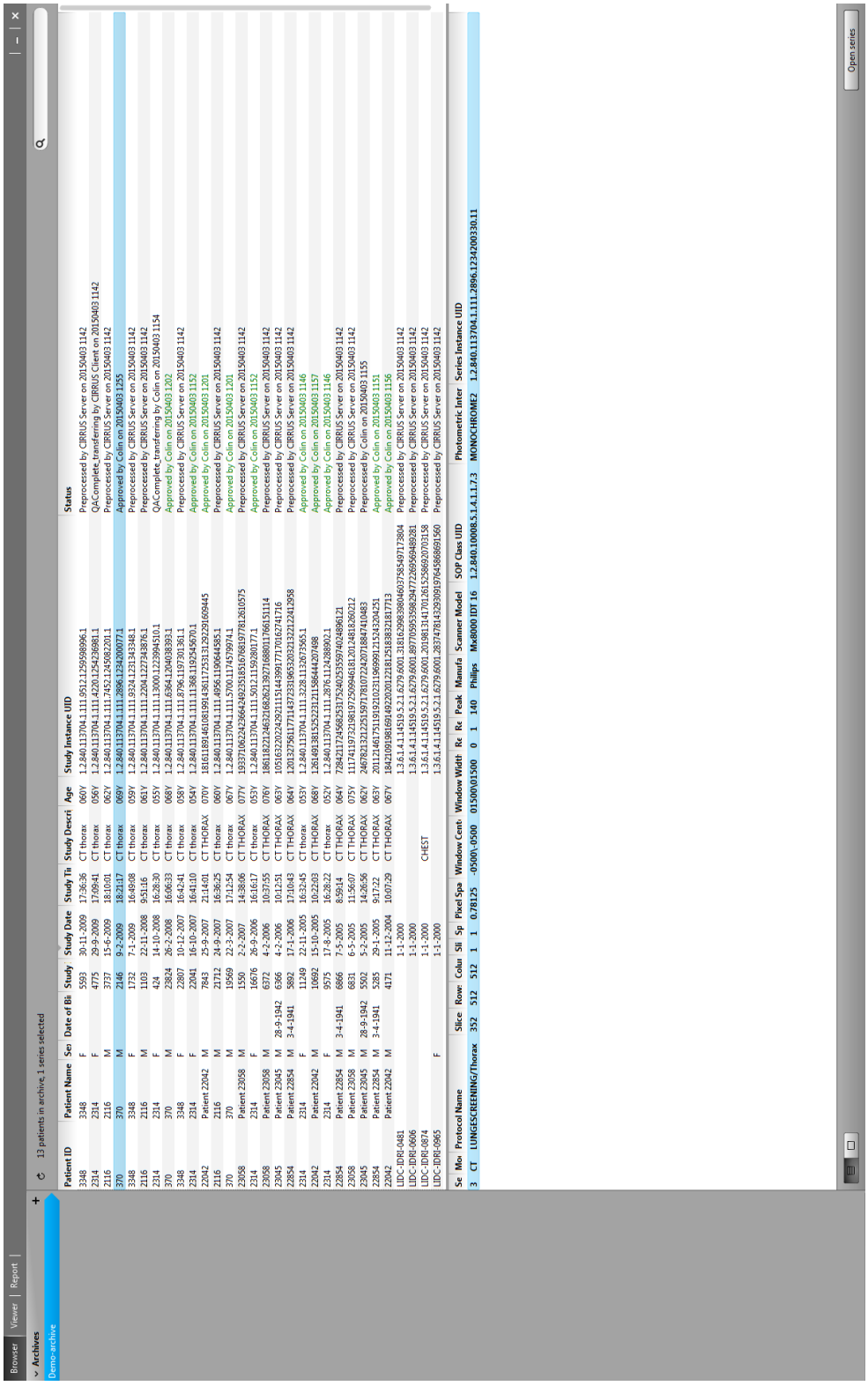


Figure 7.1: Patient browser. The order of subjects provides a worklist.

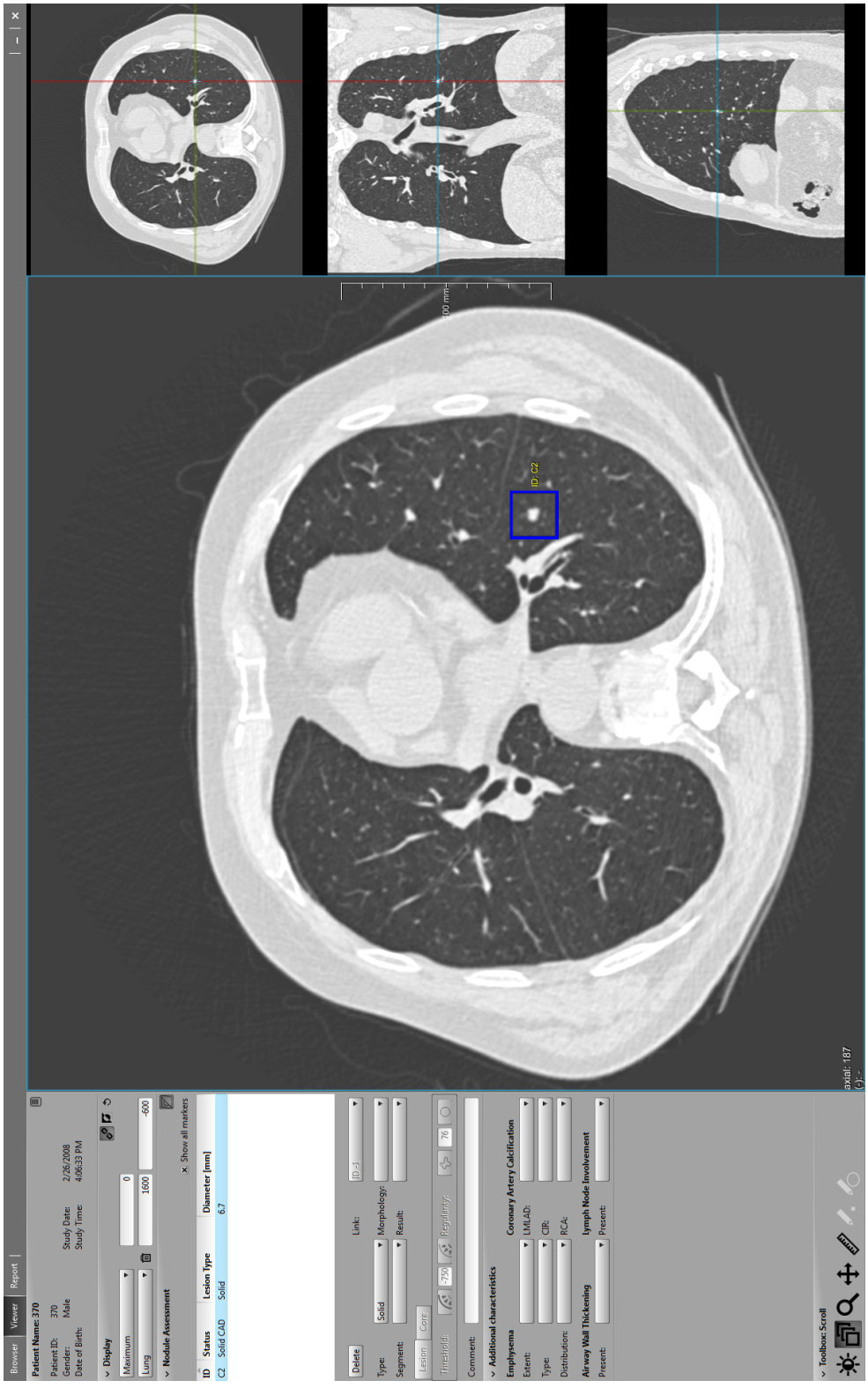


Figure 7.2: Example of the detection of a new nodule by CAD. Image shows primary screen of workstation, showing the new CT scan acquired in February, 2008. See Fig. 7.3 for second screen.

Figure 7.3: Example of the detection of a new nodule by CAD. Image shows second screen of workstation, showing the prior CT scan acquired in March, 2007. The corresponding location of the new nodule is automatically displayed thanks to the integrated registration, and this shows that the nodule is not yet visible in this scan. See Fig. 7.2 for primary screen.

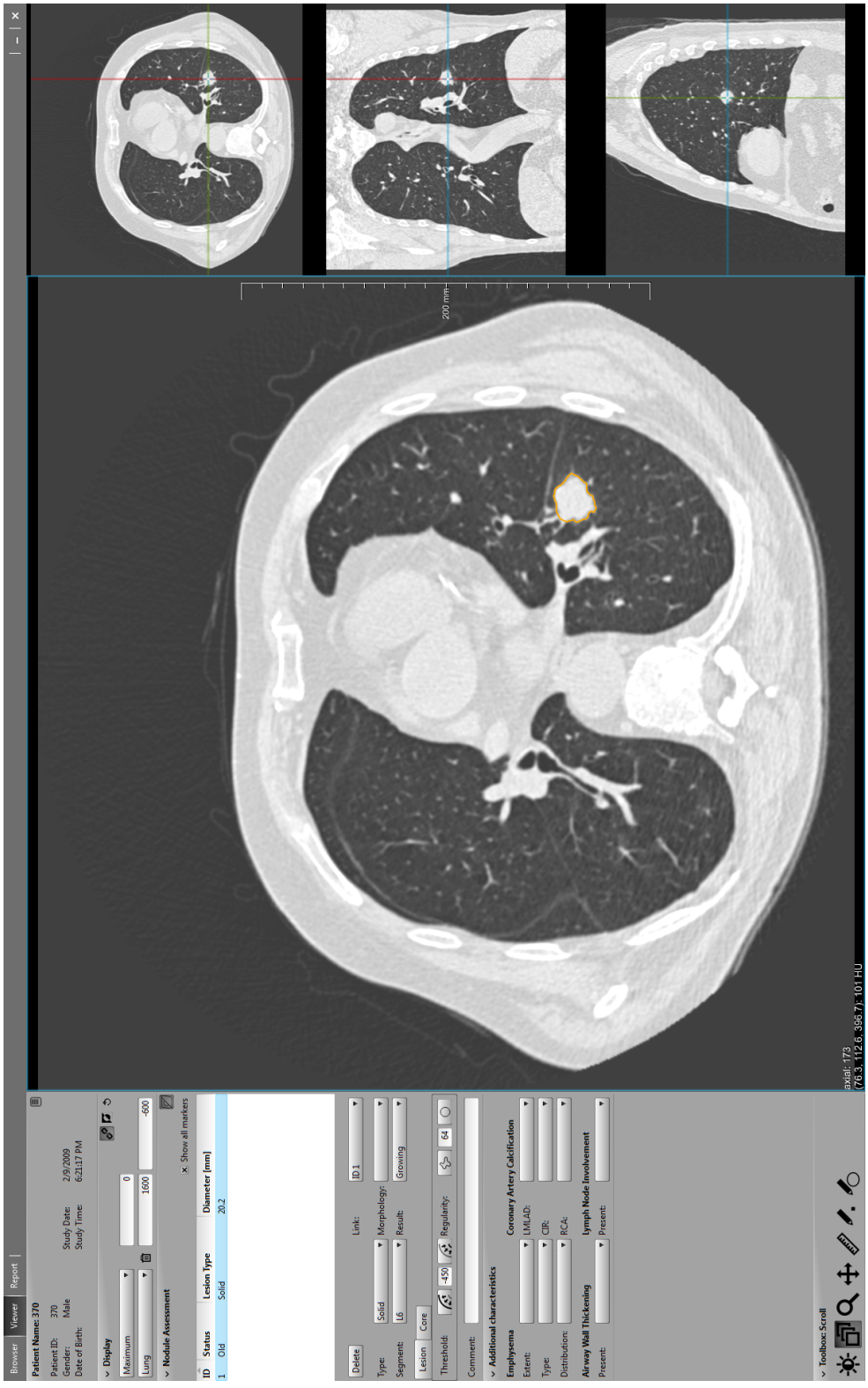


Figure 7.4: Example of a lung cancer found during screening. This is the same subjects as shown in Fig. 7.2 and Fig. 7.3. Subject was referred to pulmonologist. Diagnosis: Squamous cell carcinoma, stage T1N0M0. Image shows primary screen of workstation. See Fig. 7.5 for second screen.

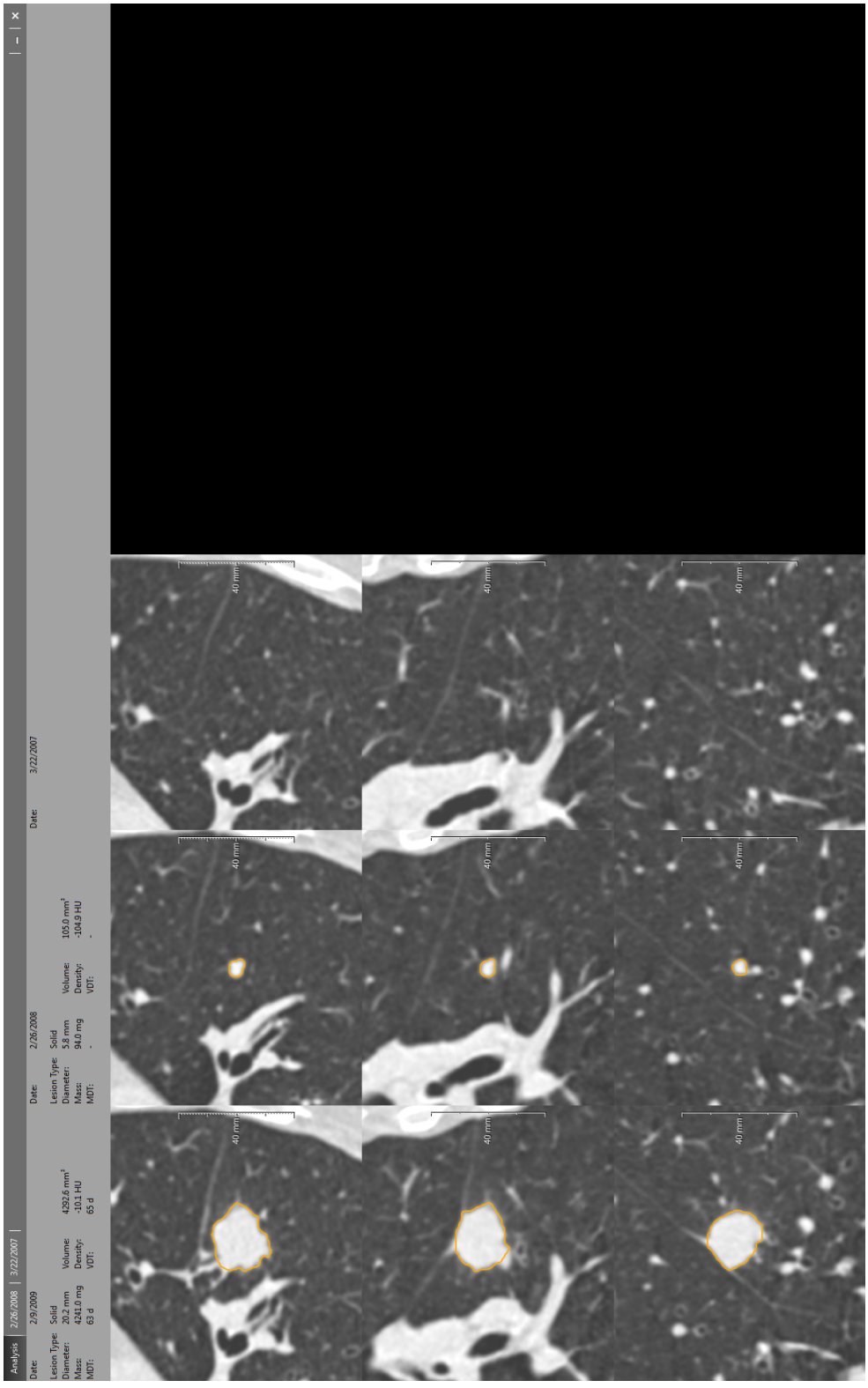


Figure 7.5: Example of a lung cancer found during screening. The Analysis tab shows the temporal evolution of this fast growing cancer. From right to left, the region without nodule in 2007, a 5.8 mm nodule in 2008 and a cancer of more than 2 centimeter but still early stage, in 2009. Image shows second screen of workstation. See Fig. 7.4 for primary screen.

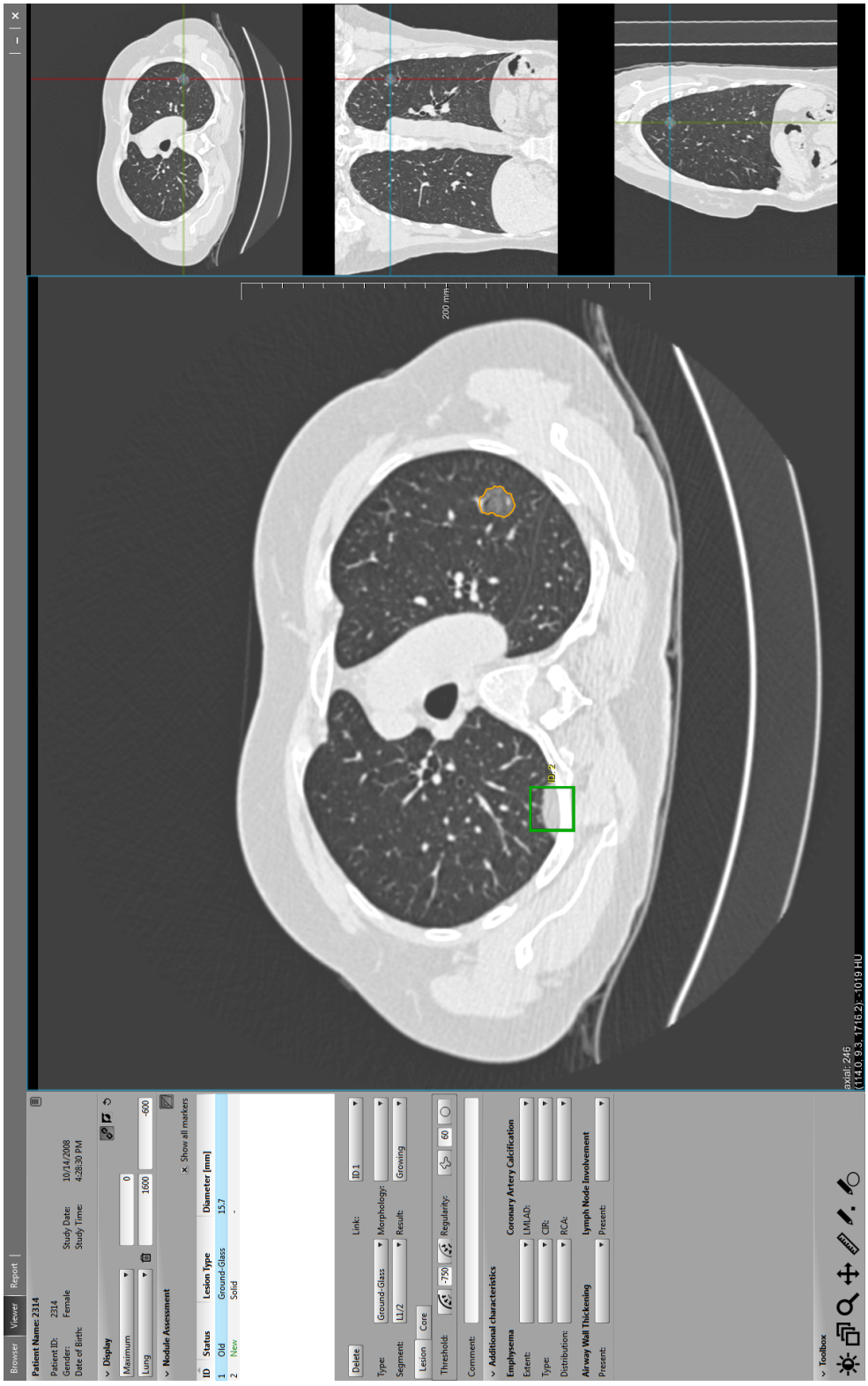


Figure 7.6: Example of a slowly growing non-solid lesion. Image shows primary screen of workstation. See Fig. 7.7 for second screen.

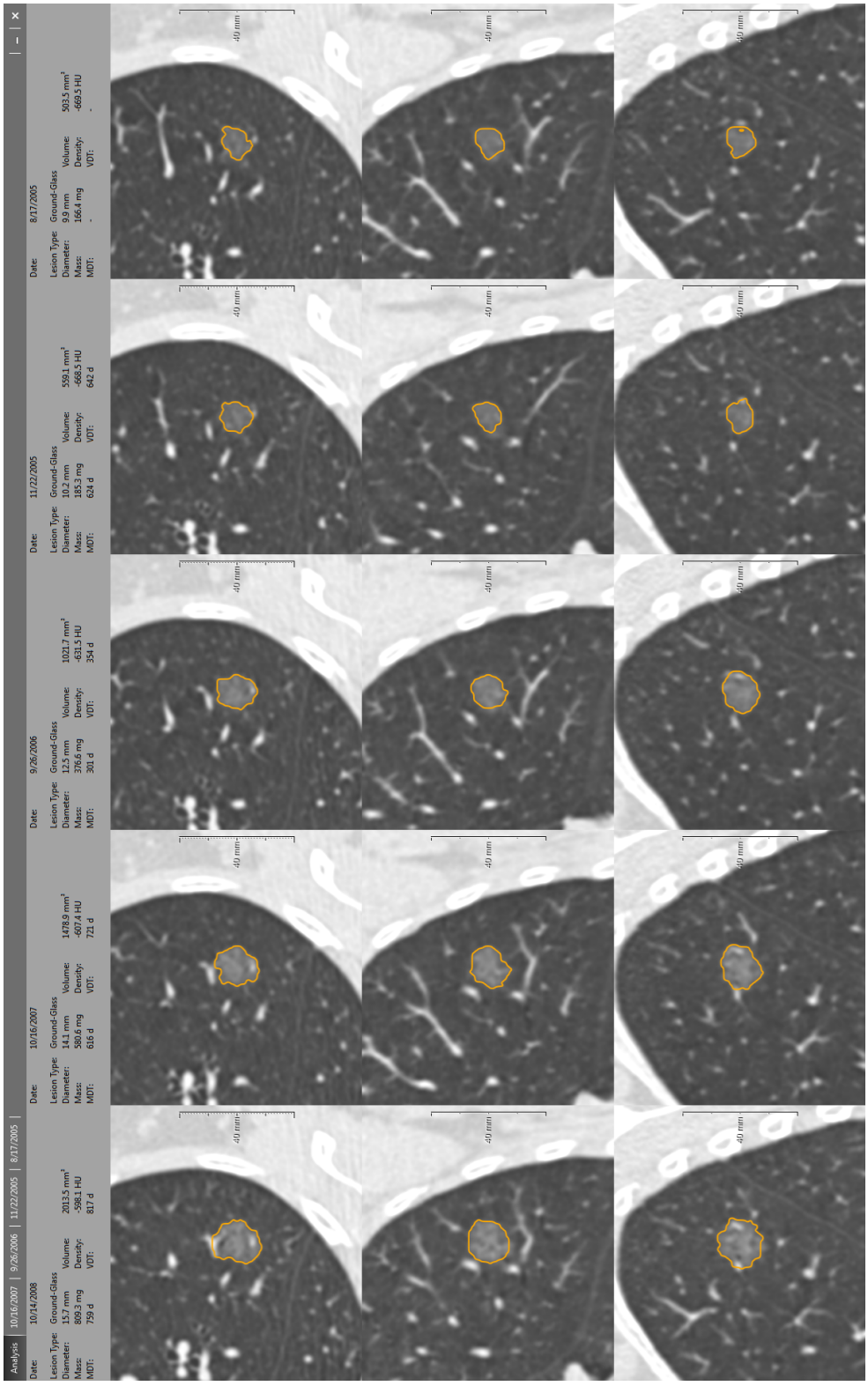


Figure 7.7: Example of a slowly growing non-solid lesion. Image shows second screen of workstation. In the course of three years, the volume of the nodule increased four-fold. See Fig. 7.6 for primary screen.

Figure 7.8: Example of the development of a solid core in a subsolid lesion. Subject was referred to pulmonologist. Diagnosis: adenocarcinoma, stage T1N0M0. Image shows primary screen of workstation. See Fig. 7.9 for second screen.

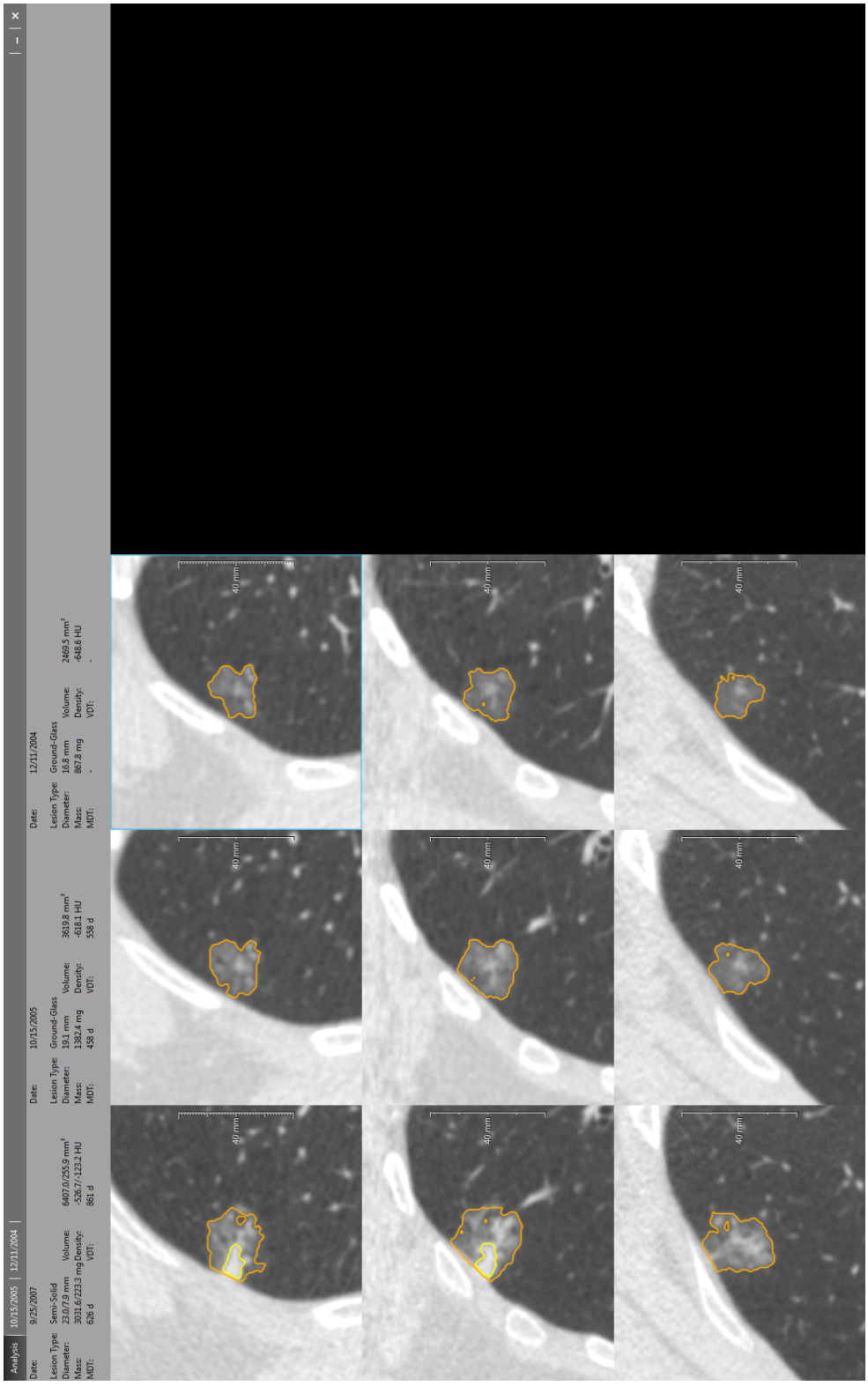


Figure 7.9: Example of the development of a solid core in a subsolid lesion. In the Analysis tab the appearance of a solid core is evident. A segmentation of the solid core is included in the current scan (yellow contour). Image shows second screen of workstation. See Fig. 7.8 for primary screen.

Chest Screening Report

Name	Patient 22042
Sex	M
Age	70
ID	22042
Visit	3
CT Scan Date	9/25/2007
Signed off by	Colin
Comments	
LungRADS Assessment	Category 4B based on nodule ID 1. Management: chest CT with or without contrast, PET/CT and/or tissue sampling depending on the probability of malignancy and comorbidities. PET/CT may be used when there is a > = 8 mm solid component.

		Finding	Baseline 12/11/2004	Visit 2 10/15/2005	Visit 3 9/25/2007	Comments
1		Location	Segment L4, Slice 336	Segment L4, Slice 323	Segment L4, Slice 350	
		Status	New	Old	Old	
		Result		Growing	Growing	
		Type	Ground-Glass	Ground-Glass	Semi-Solid	
		Equivalent Diameter	16.8 mm	19.1 mm	23.0 / 9.7 mm	
		Mass	867.8 mg	1382.4 mg	3031.6 / 405.2 mg	
		Axis long/short	21.0 / 18.7 mm	25.0 / 18.7 mm	28.2 / 20.4 mm	
		Description				
		Volume doubling time		558 d	862 d	
		Mass doubling time		459 d	627 d	
		Malignancy probability	24.04%	30.75%	48.75%	
2		Location	Segment L4, Slice 360	Segment L4, Slice 345	Segment L4, Slice 367	
		Status	New	Old	Old	
		Result		Unchanged	Smaller	
		Type	Ground-Glass	Ground-Glass	Ground-Glass	
		Equivalent Diameter	7.6 mm	7.5 mm	6.4 mm	
		Mass	94.2 mg	88.5 mg	62.8 mg	
		Axis long/short	9.5 / 6.5 mm	9.1 / 6.7 mm	7.7 / 7.1 mm	
		Description				
		Volume doubling time		-2953 d	-1111 d	
		Mass doubling time		-3420 d	-1435 d	
		Malignancy probability	4.96%	4.45%	2.95%	
3		Location	Segment L3, Slice 181	Segment L3, Slice 184	Segment L3, Slice 217	
		Status	New	Old	Old	
		Result		Growing	Smaller	
		Type	Calcified	Calcified	Calcified	
		Equivalent Diameter	2.7 mm	3.0 mm	2.6 mm	
		Mass	13.6 mg	18.3 mg	11.4 mg	
		Axis long/short	3.8 / 2.7 mm	3.9 / 3.2 mm	2.6 / 2.6 mm	
		Description				
		Volume doubling time		778 d	-1097 d	
		Mass doubling time		719 d	-1040 d	

Figure 7.10: Example of the reports. This example includes a nodule which developed a solid core. This is the same subjects as shown in Fig. 7.8 and Fig. 7.9. A nodule with a solid core larger than ≥8 mm falls into LungRADS category 4B and needs further follow-up, as stated in this report. Image shows primary screen of workstation. See Fig. 7.11 for second screen.

Figure 7.11: Example of the reports. Image shows second screen of workstation, which shows the prior reports of the same subject. See Fig. 7.10 for primary screen.

Summary

The preceding chapters of this thesis have described different methods for the detection and characterization of pulmonary nodules in thoracic CT scans. In this chapter, we provide a general summary of this thesis and briefly describe the results of each chapter.

Lung cancer is the most deadly cancer in both men and women. The American Cancer Society estimates that lung cancer will account for 27% of all cancer-related deaths in the United States in 2015[1]. This can be largely attributed to the fact that at present, the 5-year survival rate for all stages combined is only 17%[1]. The 5-year survival rate is 54% for cases detected when the disease is still localized, but only 15% of lung cancers are diagnosed at this early stage[1]. Therefore, early detection of lung cancer, in which it is still treatable, is of major importance to reduce lung cancer mortality.

Early stage lung cancer manifests itself as pulmonary nodules, which are described as round opacities, well or poorly defined, measuring up to 3 cm in diameter[12]. Thin-slice helical chest CT scans have a sub-millimeter resolution at which small pulmonary nodules can be detected[13]. Computer-aided detection of lung nodules has the potential to increase reader sensitivity for the detection of pulmonary nodules and may reduce reading time. Furthermore, automated characterization of pulmonary nodules may assist the radiologist in assessing the likelihood of malignancy of lung nodules.

Chapter 2 described a novel detection system for subsolid nodules. Detection of subsolid nodules by radiologists has increased due to the use of thin-slice CT and the implementation of lung cancer screening trials and as a consequence, their prevalence and malignancy rate are better understood. Subsolid nodules are less common, but have a higher malignancy rate than solid nodules[20]. Most published work on CAD for lung nodules primarily focused on solid nodules and therefore, we proposed a novel subsolid CAD system which aims to detect subsolid nodules. A combination of intensity, shape, texture, and context features was used to discriminate subsolid nodules from false positive candidates. The CAD system was trained with all subsolid nodules from one site and validated with subsolid nodules from another site of the NELSON lung cancer screening trial. In total, 182 subsolid nodules were included. The system achieved a sensitivity of 80% at an average of 1.0 false positives per scan.

The detection and quantification of micronodules was described in Chapter 3. Presence of disseminated micronodules on thoracic CT is the radiological sign of silicosis. In a study by Mets et al.[76], the authors reported that subjects having more than 13 micronodules should be considered at high risk for silicosis. Identification and quantification of micronodule load is tedious for human observers. We presented a

system for automatic detection of micronodules. A template matching approach is used to identify potential micronodule locations. Then, a supervised classification scheme is used to filter out false positive candidates. The system was validated on an independent data set of 54 CT scans in which four risk categories were defined. The CAD system correctly classified 83% of all subjects correctly, and obtained a weighted kappa of 0.76. This system can be used to automatically classify scans as low, intermediate, high risk, or manifest silicosis. This paves the way for automated risk assessment for silicosis and the system could be included in a lung cancer screening program.

In Chapter 4, the largest publicly available reference database for pulmonary nodules was used to benchmark the performance of state-of-the-art CAD systems. The database contained 888 CT scans from a variety of vendors and scanner models. Three CAD systems were evaluated: one academic system and two commercial systems. We showed that the commercial prototype system, *Herakles*, was able to detect the vast majority of nodules at a relatively low false positive rate. The performance of the academic system and the other commercial system varied substantially for different data sources. Furthermore, our observer experiment with four experienced radiologists showed that *Herakles* was able to find nodules missed in the extensive annotation process used to annotate the public database. We published our results on a public website, and facilitate a framework where other researchers can upload their results and obtain exactly the same FROC curves as presented in this chapter. This study contributed to a better reference standard of the public data set, and made a comparison of the performance of many CAD systems on this data set possible.

Automatic classification of pulmonary nodules into solid, part-solid, and non-solid was described in Chapter 5. We randomly selected 50 solid, 50 part-solid, and 50 non-solid nodules from a large lung cancer screening trial. An automatic segmentation-based classification method was developed which included intensity, texture and segmentation-based features. In an observer experiment, we asked three independent experienced observers to classify all nodules. Performance of CAD was evaluated by comparing the agreement between CAD and human readers and among human experts using Cohen κ statistics. Pairwise agreement between CAD and human readers had a κ range between 0.54 and 0.72. The interobserver agreement among human readers ranged between 0.56 and 0.81. Thus, agreement between CAD and human readers was in the same range as among human readers. The κ values reported in this chapter indicate that there exists a substantial interobserver variability in discriminating between solid, part-solid, and non-solid nodules. CAD may aid (unexperienced) radiologists in selecting the appropriate nodule type and hence the appropriate workup for pulmonary nodules.

Chapter 6 described a system to detect interval change between consecutive CT scans. This is the only chapter which described a method that analyzes more than one scan, and generalized the detection task from nodules to any interval change. Detection of interval change is crucial in lung cancer screening, but manual comparison of CT scans is tedious. In this study, we developed an automatic system for detection of interval change. A non-rigid registration algorithm dedicated for lung CT was employed to register a current and prior scan. Using the registration, the prior image is deformed and a subtraction image is obtained by subtracting the prior from the current scan. A CAD system which identifies potential interval change on the subtraction images and subsequently calculates a set of intensity, shape, and context features was developed to distinguish true change from subtraction artifacts. The system was validated on a set of 174 scan pairs from a large lung cancer screening trial. FROC analysis showed that the system was able to detect 70% of all annotated change at an average of 2.0 false positives per scan. The quality of the subtraction images was rated high by an experienced radiologist. This chapter indicated that automatic detection of interval change is feasible and that subtraction may be a useful tool to identify interval change in a lung cancer screening setting.

Finally, the integration of the presented methods into an optimized reading workstation was discussed in Chapter 7. The translation from research algorithms to commercial software is not trivial and this chapter explains the steps of this process. Our solution for high throughput lung cancer screening was described in detail.

General discussion

In this thesis, we focused on the detection and characterization of pulmonary nodules in thoracic CT scans. Our chapter on automatic detection of interval change between consecutive CT scans was initiated with the idea to detect new and/or growing nodules but was generalized into an automatic system to detect any interval change between consecutive CT scans. This chapter provides a general discussion about the contributions of this thesis and suggestions for future work are given.

CAD for lung nodules in chest CT has been around for more than two decades now, but with the increasing amount and quality of data used to train CAD algorithms and the continuing development of novel image analysis algorithms, performance is still increasing. However, CAD performance is still inferior to the performance of radiologists.

The appearance of lung nodules on chest CT scans can be very different and therefore, "one system to rule them all" may be suboptimal. Combination of dedicated algorithms for detection of certain subtypes, such as the dedicated CAD system for subsolid nodules described in Chapter 2 and the dedicated micronodule CAD system in Chapter 3, may be a good approach to achieve a good performance for all subtypes of nodules. However, each CAD system generates false positives at a certain rate and hence, the amount of false positives may quickly add up when many CAD systems are combined. This is one of the reasons why effective combination of different CAD systems is not trivial. Effective combination of CAD systems is not covered in this thesis and needs further investigation.

Further improvements for the performance of nodule detection systems may be expected when the size of the training datasets is increased. As predicted by Gordon Moore, co-founder of Intel, already in 1965, computing power is increasing at an exponential pace. This insight is nowadays generally known as Moore's law. Because computers become more powerful, CAD algorithms can be trained faster and faster and more and more data can be fed into the training process. Hence, a straightforward step to improve performance of the presented CAD systems is to extend the training data. Nowadays, more and more data from trials are made publicly available. The NLST trial is a good example where CT and annotation data can be requested and this can easily provide a research group with more than 10,000 CT scans. Storage space is getting cheaper and cheaper and hence, more and more data can be stored for a longer period of time. These developments make it possible to build up a tremendous amount of training data that CAD performance will most likely benefit from. In addition, the diversity of the training data is another important aspect to ensure that CAD systems perform robustly on all types of data. As shown in Chapter 4 of this thesis, the performance of a system which is trained with only one type of data (ISICAD was only trained with NELSON data) detoriates

when the CAD system is confronted with other types of data.

A hot topic which is of interest to further improve the performance of nodule CAD is deep learning. Deep learning is a term which refers to algorithms that attempt to automatically learn representations of the data. In most CAD systems until now, the features which are used by the supervised classifier are handcrafted by the developers, or chosen from a large set of handcrafted features described in the literature. In deep learning, the handcrafted features are replaced by efficient algorithms for feature learning, essentially learning which features are best to describe the underlying data. Since the process of learning the representation of the underlying data is very computationally expensive, these methods have only started to outperform traditional methods in the last couple of years, when computing power (for example through the use of GPUs) was strong enough to facilitate this. We recently started to explore the applicability of such approaches to nodule detection[114].

Another area where we can improve is the detection of rare nodules, such as endobronchial nodules, hilar lesions, or carcinomas presenting as irregular wall thickening of lung bullae (cysts). A major problem with these lesions is the small amount of training data available to train a CAD system. However, these lesions represent an important part of the interval carcinomas in a lung cancer screening setting[115] and our recent evaluation studies indicate that cancers missed by our current CAD systems are often from one of these groups. Therefore, efforts need to be made to develop CAD systems focusing on these types of nodules. In Europe, many lung cancer screening trials have been running for the last decade and consortia are being formed to exchange data. Retrospective analysis of all CT scans may lead to enough training data for these types of nodules. In addition, a first version of a CAD system may be used to search large amounts of not annotated data to gather more training data, similar to the method we used in Chapter 3 to collect micronodule examples.

All presented algorithms, except for the interval change detection system (Chapter 6), focus on the analysis of a single CT scan. In a lung cancer screening setting where subjects will be screened between the age of 55 and 77 years old, there will practically always be a prior scan available. Results of prior analyses may have been accepted/approved by a radiologist and this information can be used to guide CAD systems applied to the new scan. A simple example is the propagation of annotated nodules as CAD marks on a new scan, which subsequently need to be checked for persistence and/or growth. Future development of CAD algorithms should include prior scans in the analysis. Our proposed interval change detection system described in Chapter 6 is a first attempt in that direction. The non-rigid registration method described in this chapter can be used to align current and prior images. Another example is an algorithm which estimates the probability of malignancy of CAD marks

based on intensity, shape, texture on both scans and also growth and density change can be taken into account. If we follow this direction, we basically go from analyzing scans to analyzing subjects, or, to say it differently, from analyzing nodules as they appear on one time point to analyzing nodules as they evolve over time.

Computer models/algorithms may be of value in the characterization and management of nodules, but this topic is not extensively covered in this thesis. Our chapter on nodule classification into solid, part-solid, non-solid is only a start into this direction and follows clear visual criteria used by radiologists in the management of nodules. Possibly, direct estimation of malignancy probability by computer algorithms based on intensity, growth, texture, etc is feasible in the future. This is an interesting area of future research. Potentially, if malignancy probability can be accurately predicted, this may even decrease the false positive rate of screening tests, which is one of the biggest concerns around lung cancer screening.

There are still several concerns around the use of lung nodule CAD in clinical practice or in a screening scenario. First of all, the performance of CAD is suboptimal and hence important nodules are sometimes missed. CAD in combination with a technician or a trained reader is an interesting combination which may perform similar to a radiologist or may be used to filter out negative screening scans. In the latter scenario, only the difficult cases would need to be inspected by a radiologist. However, the validation of this paradigm is not covered in this thesis and needs further investigation. Secondly, the integration of CAD systems into the clinical and screening workflow is not trivial. In our last chapter, we have described the aspects which are important when the proposed methods of this thesis are integrated into a workstation. Although we described how to efficiently integrate the proposed methods into a screening workstation, integration into other environments, such as an oncology workstation, PACS systems, etc. can be accomplished in many different ways. Finally, history teaches us that the use of CAD in screening leads to debate. CAD has been used as a second reader in breast cancer screening for a longer period now and studies which investigate the benefits of CAD report mixed results. Gilbert et al. found that single reading with CAD is as effective as double reading by two radiologists[116]. However, more recently, Fenton et al. reported that CAD is associated with decreased specificity and not with improvement in the detection rate or prognostic characteristics of invasive breast cancer[117]. The same debate may be expected when CAD will be actively used in lung cancer screening. Compared to breast cancer screening, the use of CAD in lung cancer screening may be more imperative because of the labor-intensive nature of reading three-dimensional screening CT scans.

Samenvatting

Longkanker is de meest dodelijke vorm van kanker bij zowel mannen als vrouwen. In Nederland sterven elk jaar meer dan 10.000 mensen aan longkanker. De prognose bij longkanker is slecht. Vijf jaar na het stellen van de diagnose is slechts 17% van de mensen met longkanker nog in leven. Dit komt hoofdzakelijk doordat symptomen over het algemeen pas ontstaan wanneer de ziekte zich al in een laat stadium bevindt. Wanneer de ziekte in een vroeg stadium wordt ontdekt, is de vijfjaarsoverlevingskans veel groter: 54%. Om de sterfte door longkanker terug te dringen is het dus cruciaal om longkanker in een vroeg stadium te ontdekken. Helaas wordt maar 15% van de longkankers in dit stadium ontdekt.

Pulmonale nodulen zijn radiografische bevindingen die kunnen duiden op een vroeg stadium van longkanker. Een pulmonale nodule is een ronde afwijking in de longen van maximaal 3 cm in diameter, ook wel bekend als het 'vlekje op de long'. Het opsporen en vaststellen van de maligniteit van pulmonale nodulen is essentieel om longkanker in een vroeg stadium te diagnosticeren. CT scans van de thorax hebben een resolutie waarmee pulmonale nodulen goed op te sporen zijn. Computergestuurde detectie (CAD) systemen die automatisch pulmonale nodulen in thorax CT scans opsporen hebben veel potentie. Ze kunnen de sensitiveit van radiologen voor het opsporen van nodulen verbeteren en de interpretatietijd verkorten. Verder zou automatische karakterisatie van pulmonale nodulen de radioloog kunnen helpen bij het bepalen van de aard (kwaadaardig / goedaardig) van de nodule.

Hoofdstuk 2 beschrijft een nieuw systeem voor het opsporen van subsolide nodulen. Door de verbeterde resolutie van hedendaagse CT scans kan men dit type nodulen tegenwoordig goed opsporen. Verder hebben de longkanker screening studies veel CT data gegenereerd die bestudeerd kunnen worden. Hierdoor begrijpen we de prevalentie en maligniteit van subsolide nodulen beter. Subsolide nodulen komen minder voor, maar hebben een grotere kans maligne te zijn. Het meeste werk dat gepubliceerd is over automatische detectie van pulmonale nodulen heeft zich gefocust op solide nodulen en daarom hebben wij een systeem ontwikkeld dat zich specifiek op de automatische detectie van subsolide nodulen richt. Het CAD systeem is getraind met alle subsolide nodulen die gevonden zijn in een van de deelnemende ziekenhuizen van de NELSON studie, en geëvalueerd op alle subsolide nodulen van een andere deelnemend ziekenhuis van de NELSON studie. In totaal zijn er 182 subsolide nodulen geïncludeerd in deze studie. Het systeem haalt een sensitiviteit van 80% bij een gemiddelde van 1 foutpositieve detectie per scan. Verder hebben we laten zien dat het systeem subsolide nodulen vindt die niet geannoteerd zijn door de radiologen in de screening studie.

De detectie en kwantificatie van micronodulen is beschreven in *hoofdstuk 3*. De

aanwezigheid van verspreide micronodulen op een thorax CT scan is het radiologische kenmerk van silicose (stoflongen). Een studie door Mets et al. rapporteerde dat mensen met meer dan 13 micronodulen een verhoogd risico hebben om silicose te ontwikkelen. Het opsporen en kwantificeren van micronodulen is erg tijdrovend voor lezers. In dit hoofdstuk presenteren we daarom een automatisch systeem voor het detecteren van micronodulen. We gebruiken een template matching methode om potentiële locaties van micronodulen te vinden. Vervolgens gebruiken we een classificatie systeem wat getraind wordt met voorbeelden om foutpositieve kandidaten eruit te filteren. Het systeem is gevalideerd op een onafhankelijke data set van 54 CT scans die in vier verschillende risicocategorieën waren onderverdeeld. Het CAD systeem wist 83% van alle deelnemers in de juiste categorie te plaatsen. De gewogen kappa waarde was 0.76. Dit systeem kan gebruikt worden voor geautomatiseerde classificatie van CT scans in laag, gemiddeld, of hoog risico op silicosis, of zichtbare silicosis. Dit systeem opent de weg naar automatische inschatting van het risico op silicosis en zou mogelijk gebruikt kunnen worden in een longkanker screeningsprogramma.

In *hoofdstuk 4* hebben we de grootste publieke referentiedatabase van pulmonale nodulen gebruikt om verschillende state-of-the-art CAD systemen te vergelijken. De database bestaat uit 888 CT scans die gemaakt zijn door verschillende scanners van verschillende fabrikanten. We hebben in deze studie 3 CAD systemen vergeleken: een CAD systeem ontwikkeld door een universiteitsgroep, en twee commerciële CAD systemen. We hebben laten zien dat het nieuwste commerciële prototype, *Herakles*, het beste presteert. De resultaten van de andere twee systemen variëren sterk voor verschillende typen data. Verder hebben we in een observer experiment laten zien dat *Herakles* nodulen kan vinden die waren gemist tijdens het uitgebreide annotatieproces dat gebruikt is om de publieke database op te bouwen. We hebben onze resultaten op een website gepubliceerd, en via deze website kunnen andere onderzoekers hun resultaten uploaden en vergelijken met onze resultaten. Op deze manier hebben we bijgedragen aan een betere referentiestandaard voor de publieke database en faciliteren we het vergelijken van CAD systemen op een publieke database.

Het automatisch classificeren van nodulen in solide, part-solide, en niet-solide nodulen is beschreven in *hoofdstuk 5*. Voor deze studie hebben we willekeurig 50 solide, 50 part-solide, en 50 niet-solide nodulen geselecteerd uit de data van twee deelnemende centra van de NELSON studie. We hebben een systeem ontwikkeld dat de intensiteit, textuur en de segmentatiekarakteristieken van een nodule gebruikt om deze te classificeren. In een observer experiment hebben we drie radiologen voor dezelfde taak gezet. We hebben vervolgens gekeken naar de overeenstem-

ming tussen de verschillende radiologen, en de overeenstemming tussen radiologen en het CAD systeem gebruikmakend van Cohen's kappa statistiek. De kappa waarden tussen CAD en de radiologen varieerde tussen 0.54 en 0.72. De kappa waarden tussen de radiologen onderling varieerde tussen 0.56 en 0.81. We kunnen dus concluderen dat de kappa waarden tussen radiologen in hetzelfde gebied liggen als tussen radiologen en CAD. Mogelijk kan CAD dus (onervaren) lezers helpen bij het bepalen van het type nodule, en daarmee helpen bij het bepalen van de juiste follow-up.

Hoofdstuk 6 beschrijft een systeem waarmee veranderingen in de longen tussen opeenvolgende CT scans kunnen worden opgespoord. Het systeem richt zich niet alleen op nodulaire veranderingen, maar elke verandering die tussen twee opeenvolgende CT scans kan worden opgespoord. Het opsporen van veranderingen tussen CT scans is cruciaal in longkanker screening, maar het visueel vergelijken van twee CT scans is erg tijdrovend. Daarom hebben we in deze studie een systeem ontwikkeld dat automatisch veranderingen opspoort. We gebruiken hiervoor een niet-rigide registratie methode die speciaal ontwikkeld is voor registreren van thorax CT scans. Met behulp van het resulterende deformatieveld vervormen we de oude scan en vervolgens trekken we die van de nieuwe scan af. Het resulterende subtractiebeeld gebruiken we vervolgens in het systeem. Het systeem spoort afwijkingen op in het subtractiebeeld en is getraind om echte veranderingen van artefacten te onderscheiden. Met behulp van FROC analyse hebben we laten zien dat het systeem 70% van alle geannoteerde veranderingen kan vinden met een gemiddelde van 2 foutpositieven per scanpaar. De kwaliteit van alle subtractiebeelden is visueel beoordeeld door een ervaren radioloog en als zeer goed beoordeeld. Deze studie toont aan dat het automatisch opsporen van veranderingen mogelijk is en dat subtractie een veelbelovende techniek kan zijn in een longkanker screening scenario.

Als laatste hebben we in *hoofdstuk 7* beschreven hoe we de systemen uit dit onderzoek hebben geïntegreerd in een screeningswerkstation. Het implementeren van systemen die ontwikkeld zijn in een onderzoeksgroep in de kliniek is niet triviaal. In het laatste hoofdstuk beschrijven we welke stappen er nodig zijn om deze systemen om te bouwen naar bruikbare, robuste systemen die in de kliniek gebruikt kunnen worden.

Publications

Papers in international journals

C. Jacobs, E.M. van Rikxoort, K. Murphy, M. Prokop, C.M. Schaefer-Prokop and B. van Ginneken. "Computer-aided detection of pulmonary nodules: a comparative study using the public LIDC/IDRI database". *European Radiology*, 2015.

A.A.A. Setio, **C. Jacobs**, J. Gelderblom and B. van Ginneken. "Automatic detection of large pulmonary solid nodules in thoracic CT images", *Medical Physics*, 42:5642-5653, 2015.

F. Ciompi, **C. Jacobs**, E.T. Scholten, M.M. Wille, P.A. de Jong, M. Prokop and B. van Ginneken. "Bag of frequencies: a descriptor of pulmonary nodules in Computed Tomography images". *IEEE Transactions on Medical Imaging*, 34:1-12, 2015.

C. Jacobs, E.M. van Rikxoort, E.T. Scholten, P.A. de Jong, M. Prokop, C.M. Schaefer-Prokop and B. van Ginneken. "Solid, Part-Solid, or Non-solid?: Classification of Pulmonary Nodules in Low-Dose Chest Computed Tomography by a Computer-Aided Diagnosis System". *Investigative Radiology*, 50:168-173, 2015.

B.C. Lassen, **C. Jacobs**, J.M. Kuhnigk, B. van Ginneken and E.M. van Rikxoort. "Robust semi-automatic segmentation of pulmonary subsolid nodules in chest computed tomography scans". *Physics in Medicine and Biology*, 60:1307-1323, 2015.

S.J. van Riel, C.I. Sánchez, A.A. Bankier, D.P. Naidich, J. Verschakelen, E.T. Scholten, P.A. de Jong, **C. Jacobs**, E.M. van Rikxoort, L. Peters-Bax, M. Snoeren, M. Prokop, B. van Ginneken and C.M. Schaefer-Prokop. "Observer Variability for Classification of Pulmonary Nodules on Low-Dose CT Images and Its Effect on Nodule Management", *Radiology*.

E.T. Scholten, P.A. de Jong, B. de Hoop, R. van Klaveren, S. van Amelsvoort-van de Vorst, M. Oudkerk, R. Vliegenthart, H.J. de Koning, C.M. van der Aalst, R.M. Vernhout, H.J.M. Groen, J.-W.J. Lammers, B. van Ginneken, **C. Jacobs**, W.P.T.M. Mali, N. Horeweg, C. Weenink, E. Thunnissen, M. Prokop and H.A. Gietema. "Towards a close computed tomography monitoring approach for screen detected subsolid pulmonary nodules?". *European Respiratory Journal*, 45:765-773, 2015.

E.T. Scholten, P.A. de Jong, **C. Jacobs**, B. van Ginneken, S.J. van Riel, M.J. Willemink, R. Vliegenthart, M. Oudkerk, H.J. de Koning, N. Horeweg, M. Prokop, W.P.T.M. Mali and H.A. Gietema. "Interscan variation of semi-automated volumetry of subsolid pulmonary nodules". *European Radiology*, 25:1040-1047, 2015.

E.T. Scholten, **C. Jacobs**, B. van Ginneken, S.J. van Riel, R. Vliegenthart, M. Oudkerk, H.J. de Koning, N. Horeweg, M. Prokop, H.A. Gietema, W.P.T.M. Mali and P.A. de Jong. "Detection and quantification of the solid component in pulmonary subsolid nodules by semiautomatic segmentation". *European Radiology*, 25:488-496, 2015.

M.M. Winkler Wille, S.J. van Riel, Z. Saghir, A. Dirksen, J.H. Pedersen, **C. Jacobs**, L.H. Thomsen, E.T. Scholten, L.T. Skovgaard and B. van Ginneken. "Predictive Accuracy of the PanCan Lung Cancer Risk Prediction Model - External Validation based on CT from the Danish Lung Cancer Screening Trial". *European Radiology*, 2015.

C. Jacobs, E.M. van Rikxoort, T. Twellmann, E.T. Scholten, P.A. de Jong, J.M. Kuhnigk, M. Oudkerk, H.J. de Koning, M. Prokop, C.M. Schaefer-Prokop and B. van Ginneken. "Automatic Detection of Subsolid Pulmonary Nodules in Thoracic Computed Tomography Images". *Medical Image Analysis*, 18:374-384, 2014.

E.T. Scholten, B. de Hoop, **C. Jacobs**, S. van Amelsvoort-van de Vorst, R.J. van Klaveren, M. Oudkerk, R. Vliegenthart, H.J. de Koning, C.M. van der Aalst, W.T.M. Mali, H.A. Gietema, M. Prokop, B. van Ginneken and P.A. de Jong. "Semi-automatic quantification of subsolid pulmonary nodules: comparison with manual measurements". *PLoS One*, 8:e80249, 2013.

E.T. Scholten, **C. Jacobs**, B. van Ginneken, M.J. Willemink, J.M. Kuhnigk, P.M.A. van Ooijen, M. Oudkerk, W.P.T.M. Mali and P.A. de Jong. "Computer-Aided Segmentation and Volumetry of Artificial Ground-Glass Nodules at Chest CT". *American Journal of Roentgenology*, 201:295-300, 2013.

O.M. Mets, R.A. van Hulst, **C. Jacobs**, B. van Ginneken and P.A. de Jong. "Normal Range of Emphysema and Air Trapping on CT in Young Men". *American Journal of Roentgenology*, 199:336-340, 2012.

Papers in conference proceedings

B. van Ginneken, A.A.A. Setio, **C. Jacobs** and F. Ciompi. "Off-the-shelf Convolutional Neural Network features for pulmonary nodule detection in computed tomography scans". In: *IEEE International Symposium on Biomedical Imaging*, pages 286-289, 2015.

A.A.A. Setio, **C. Jacobs**, F. Ciompi, S.J. van Riel, M.M.W. Wille, A. Dirksen, E.M. van Rikxoort and B. van Ginneken. "Computer-aided detection of lung cancer: combining pulmonary nodule detection systems with a tumor risk prediction model". In:

Medical Imaging, volume 9414 of Proceedings of the SPIE, page 94141O, 2015.

F. Ciompi, **C. Jacobs**, E.T. Scholten, M.M.W. Wille, M. Prokop and B. van Ginneken. "Automatic detection of spiculation of pulmonary nodules in Computed Tomography images". In: *Medical Imaging*, volume 9414 of Proceedings of the SPIE, page 941409, 2015.

C. Jacobs, S.H.T. Opdam, E.M. van Rikxoort, O.M. Mets, J. Rooyackers, P.A. de Jong, M. Prokop and B. van Ginneken. "Automated detection and quantification of micronodules in thoracic CT scans to identify subjects at risk for silicosis". In: *Medical Imaging*, volume 9035 of Proceedings of the SPIE, page 90351I, 2014.

Z. Saghir, **C. Jacobs**, B. van Ginneken, M. de Bruijne, A. Dirksen, J.H. Pedersen. "Performance of segmentation software on large longitudinal database of pulmonary nodules in the Danish Lung Cancer Screening Trial (DLCST)". *European Respiratory Journal*, volume 40 (Suppl 56), page 262, 2012.

C. Jacobs, C.I. Sánchez, S.C. Saur, T. Twellmann, P.A. de Jong and B. van Ginneken. "Computer-Aided Detection of Ground Glass Nodules in Thoracic CT images using Shape, Intensity and Context Features". In: *Medical Image Computing and Computer-Assisted Intervention*, volume 6893 of Lecture Notes in Computer Science, pages 207-214, 2011.

C. Jacobs, K. Murphy, T. Twellmann, P.A. de Jong and B. van Ginneken. "Computer-Aided Detection of Solid and Ground Glass Nodules in Thoracic CT images using two independent CAD systems". In: *The Fourth International Workshop on Pulmonary Image Analysis*, 2011.

Abstracts in conference proceedings

A. Ritchie, M. Tammemagi, **C. Jacobs**, W. Zhang, J. Mayo, H. Roberts, M. Gingras, S. Pasian, L. Stewart, S. Tsai, D. Manos, J. Seely, P. Burrowes, R. Bhatia, S. Atkar-Khattra, B. van Ginneken, M. Tsao, and S. Lam, "Automated measurement of malignancy risk of lung nodule detected by screening computed tomography", In: World Conference on Lung Cancer, 2015.

A. Ritchie, C. Sanghera, **C. Jacobs**, W. Zhang, J. Mayo, H. Roberts, M. Gingras, S. Pasian, L. Stewart, S. Tsai, D. Manos, J. Seely, P. Burrowes, R. Bhatia, S. Atkar-Khattra, B. van Ginneken, M. Tammemagi and S. Lam, "Computer Vision Tool and Technician as First Reader of Lung Cancer Screening CT", In: World Conference on Lung

Cancer, 2015.

E.T. Scholten, **C. Jacobs**, B. van Ginneken, S.J. van Riel, R. Vliegenthart, M. Oudkerk, M. Prokop, H. Gietema and P.A. de Jong. "Detection and quantification of the solid component in pulmonary subsolid nodules by semiautomatic segmentation", In: European Congress of Radiology, 2015.

F. Ciompi, B. de Hoop, **C. Jacobs**, M. Prokop, P. a de Jong and B. van Ginneken. "Automatic Classification of Perifissural Pulmonary Nodules in Thoracic CT Images". In: Annual Meeting of the Radiological Society of North America, 2014.

B. van Ginneken, **C. Jacobs**, E.T. Scholten, M. Prokop and P.A. de Jong. "Feasibility of Rapid Reading of CT Lung Cancer Screening with Computer-Aided Detection Support". In: Annual Meeting of the Radiological Society of North America, 2014.

B.C. Lassen, J. Kuhnigk, **C. Jacobs**, E.M. van Rikxoort and B. van Ginneken. "Fully Automatic Volumetric Segmentation of Pulmonary Nodules: Evaluation using the Complete LIDC/IDRI Database". In: Annual Meeting of the Radiological Society of North America, 2014.

A.A.A. Setio, J. Gelderblom, **C. Jacobs** and B. van Ginneken. "Automatic Detection of Large Pulmonary Nodules in Thoracic CT Images". In: Annual Meeting of the Radiological Society of North America, 2014.

B. van Ginneken, E.M. van Rikxoort, S.J. Lafebre, **C. Jacobs**, M. Schmidt, J.M. Kuhnigk, M. Prokop, C.M. Schaefer-Prokop, J.-P. Charbonnier, L. Hogeweg, P. Maduskar, L. Gallardo-Estrella, R. Philipsen and B. Lassen. "CIRRUS Lung: an optimized workflow for quantitative image analysis of thoracic computed tomography and chest radiography for major pulmonary diseases: chronic obstructive pulmonary disease, lung cancer and tuberculosis". In: Annual Meeting of the Radiological Society of North America, 2013.

C. Jacobs, B. van Ginneken, S. Fromme, M. Prokop and E.M. van Rikxoort. "Benchmarking computer-aided detection of pulmonary nodules on the recently completed publicly available LIDC/IDRI database". In: Annual Meeting of the Radiological Society of North America, 2013.

C. Jacobs, E.M. van Rikxoort, J.M. Kuhnigk, E.T. Scholten, P.A. de Jong, C.M. Schaefer-Prokop, M. Prokop and B. van Ginneken. "Automated characterization of pulmonary nodules in thoracic CT images using a segmentation-based classification system". In: European Congress of Radiology, 2013.

C. Jacobs, E.M. van Rikxoort, J.M. Kuhnigk, E.T. Scholten, P.A. de Jong, C.M. Schaefer-Prokop, M. Prokop and B. van Ginneken. "Non-solid, part-solid or solid? Classification of pulmonary nodules in thoracic CT by radiologists and a computer-aided diagnosis system". In: European Congress of Radiology, 2013.

S.J. van Riel, E.M. van Rikxoort, **C. Jacobs**, S. Schalekamp, M. Prokop, B. van Ginneken, P.A. de Jong, E.T. Scholten, H.A. Gietema and C.M. Schaefer-Prokop. "Intra- and inter-reader variability of pulmonary nodule classification according to the Fleischner guidelines: clinical consequences". In: Annual Meeting of the Radiological Society of North America, 2013.

S.J. van Riel, C.M. Schaefer-Prokop, E.M. van Rikxoort, B. van Ginneken, M. Prokop, S. Schalekamp, **C. Jacobs**, P.A. de Jong, H.A. Gietema and E.T. Scholten. "Impact of section thickness on classification of pulmonary nodules into solid, part-solid, and non-solid: an observer study". In: Annual Meeting of the Radiological Society of North America, 2013.

C. Jacobs, J. Kuhnigk, E.M. van Rikxoort, T. Twellmann, P.A. de Jong, H. Gietema, B. Lassen, C.M. Schaefer-Prokop, M. Prokop and B. van Ginneken. "Optimized workflow for low dose thoracic CT lung cancer screening: automated detection, measurement, temporal matching and volumetry and mass analysis, individualized prediction of cancer risk, structured reporting with follow-up recommendation". In: Annual Meeting of the Radiological Society of North America, 2012.

C. Jacobs, E.M. van Rikxoort, T. Twellmann, P.A. de Jong, C.M. Schaefer-Prokop, M. Prokop and B. van Ginneken. "Improved computer-aided detection of pulmonary nodules by combining a solid and subsolid nodule CAD system". In: Annual Meeting of the Radiological Society of North America, 2012.

C. Jacobs, E.T. Scholten, S.C. Saur, T. Twellmann, P.A. de Jong and B. van Ginneken. "Computer-Aided Detection of Ground Glass Nodules in Lung Cancer Screening: Retrospective Evaluation of Potential Benefit". In: European Congress of Radiology, 2012.

B.C. Lassen, M. Schmidt, E.M. van Rikxoort, B. van Ginneken, **C. Jacobs**, J. Kuhnigk and M. Prokop. "Automated and interactive image analysis workstation for the extraction of imaging biomarkers related to chronic obstructive pulmonary disease from thoracic computed tomography scans". In: Annual Meeting of the Radiological Society of North America, 2012.

Z. Saghir, **C. Jacobs**, B. van Ginneken, M. de Bruijne, A. Dirksen, J.H. Pedersen.

"Probability of malignancy based on automatic segmentation and software measurements of nodules in the Danish lung cancer screening trial (DLCST)". In: European Respiratory Society Annual Meeting, 2012.

C. Jacobs, E.T. Scholten, S.C. Saur, T. Twellmann, P.A. de Jong and B. van Ginneken. "Computer-Aided Detection of Ground Glass Nodules in Thoracic CT Images". In: Annual Meeting of the Radiological Society of North America, 2011.

Bibliography

[1] American Cancer Society. Cancer facts and figures 2015, 2015.

[2] KWF Kankerbestrijding. Kanker in Nederland tot 2020, 2011.

[3] American Cancer Society. Cancer facts and figures 2011, 2011.

[4] Röntgen W. C. Über eine neue Art von Strahlen. *Sitzungsberichte der Physikalisch-Medicinisch Gesellschaft zu Würzburg*, pages 132–141, 1895.

[5] Hounsfield G. Computerized transverse axial scanning (tomography): Part I. Description of system. *Br J Radiol*, 46:1016–1022, 1973.

[6] Doss M., Little M. P., and Orton C. G. Point/counterpoint: Low-dose radiation is beneficial, not harmful. *Med Phys*, 41:070601, 2014.

[7] American College of Radiology and Radiological Society of North America. Patient safety: Radiation dose in x-ray and CT exams, 2015.

[8] Nagatani Y., Takahashi M., Murata K., Ikeda M., Yamashiro T., Miyara T., Koyama H., Koyama M., Sato Y., Moriya H., Noma S., Tomiyama N., Ohno Y., Murayama S., and investigators of ACTIve study group. Lung nodule detection performance in five observers on computed tomography (CT) with adaptive iterative dose reduction using three-dimensional processing (AIDR 3D) in a Japanese multicenter study: Comparison between ultra-low-dose CT and low-dose CT by receiver-operating characteristic analysis. *Eur J Radiol*, 84:1401–12, 2015.

[9] van Klaveren R. J., Oudkerk M., Prokop M., Scholten E. T., Nackaerts K., Vernhout R., van Iersel C. A., van den Bergh K. A. M., van 't Westeinde S., van der Aalst C., Thunnissen E., Xu D. M., Wang Y., Zhao Y., Gietema H. A., de Hoop B., Groen H. J. M., de Bock G. H., van Ooijen P., Weenink C., Verschakelen J., Lammers J. J., Timens W., Willebrand D., Vink A., Mali W., and de Koning H. J. Management of lung nodules detected by volume CT scanning. *N Engl J Med*, 361:2221–2229, 2009.

[10] Aberle D. R., Adams A. M., Berg C. D., Black W. C., Clapp J. D., Fagerstrom R. M., Gareen I. F., Gatsonis C., Marcus P. M., and Sicks J. D. Reduced lung-cancer mortality with low-dose computed tomographic screening. *N Engl J Med*, 365:395–409, 2011.

[11] Kauczor H.-U., Bonomo L., Gaga M., Nackaerts K., Peled N., Prokop M., Remy-Jardin M., von Stackelberg O., Sculier J.-P., and the European Society of Radiology (ESR) and the European Respiratory Society (ERS). ESR/ERS white paper on lung cancer screening. *Eur Respir J*, 46: 28–39, 2015.

[12] Hansell D. M., Bankier A. A., MacMahon H., McLoud T. C., Müller N. L., and Remy J. Fleischner society: glossary of terms for thoracic imaging. *Radiology*, 246:697–722, 2008.

[13] Henschke C. I., McCauley D. I., Yankelevitz D. F., Naidich D. P., McGuinness G., Miettinen O. S., Libby D. M., Pasmantier M. W., Koizumi J., Altorki N. K., and Smith J. P. Early lung cancer action project: overall design and findings from baseline screening. *Lancet*, 354:99–105, 1999.

[14] Leader J. K., Warfel T. E., Fuhrman C. R., Golla S. K., Weissfeld J. L., Avila R. S., Turner W. D., and Zheng B. Pulmonary nodule detection with low-dose CT of the lung: agreement among radiologists. *AJR Am J Roentgenol*, 185:973–978, 2005.

[15] Seltzer S. E., Judy P. F., Adams D. F., Jacobson F. L., Stark P., Kikinis R., Swensson R. G., Hooton

S., Head B., and Feldman U. Spiral CT of the chest: comparison of cine and film-based viewing. *Radiology*, 197:73–78, 1995.

[16] Naidich D. P., Rusinek H., McGuinness G., Leitman B., McCauley D. I., and Henschke C. I. Variables affecting pulmonary nodule detection with computed tomography: evaluation with three-dimensional computer simulation. *J Thorac Imaging*, 8:291–299, 1993.

[17] Armato S. G., Roberts R. Y., Kocherginsky M., Aberle D. R., Kazerooni E. A., Macmahon H., van Beek E. J. R., Yankelevitz D., McLennan G., McNitt-Gray M. F., Meyer C. R., Reeves A. P., P.Caligiuri, Quint L. E., Sundaram B., Croft B. Y., and Clarke L. P. Assessment of radiologist performance in the detection of lung nodules: dependence on the definition of "truth". *Acad Radiol*, 16:28–38, 2009.

[18] Armato S. G., Roberts R. Y., McNitt-Gray M. F., Meyer C. R., Reeves A. P., McLennan G., Engelmann R. M., Bland P. H., Aberle D. R., Kazerooni E. A., MacMahon H., van Beek E. J. R., Yankelevitz D., Croft B. Y., and Clarke L. P. The Lung Image Database Consortium (LIDC): ensuring the integrity of expert-defined "truth". *Acad Radiol*, 14:1455–1463, 2007.

[19] Rubin G. D., Lyo J. K., Paik D. S., Sherbondy A. J., Chow L. C., Leung A. N., Mindelzun R., Schraedley-Desmond P. K., Zinck S. E., Naidich D. P., and Napel S. Pulmonary nodules on multi-detector row CT scans: Performance comparison of radiologists and computer-aided detection. *Radiology*, 234:274–283, 2005.

[20] Henschke C. I., Yankelevitz D. F., Mirtcheva R., McGuinness G., McCauley D., and Miettinen O. S. CT screening for lung cancer: Frequency and significance of part-solid and nonsolid nodules. *AJR Am J Roentgenol*, 178:1053–1057, 2002.

[21] Beyer F., Zierott L., Fallenberg E. M., Juergens K. U., Stoeckel J., Heindel W., and Wormanns D. Comparison of sensitivity and reading time for the use of computer-aided detection (CAD) of pulmonary nodules at MDCT as concurrent or second reader. *Eur Radiol*, 17:2941–2947, 2007.

[22] Li Q. Recent progress in computer-aided diagnosis of lung nodules on thin-section CT. *Comput Med Imaging Graph*, 31:248–257, 2007.

[23] Lee S. L. A., Kouzani A. Z., and Hu E. J. Automated detection of lung nodules in computed tomography images: a review. *Mach Vis Appl*, 23:151–163, 2012.

[24] Godoy M. C. B., Kim T. J., White C. S., Bogoni L., de Groot P., Florin C., Obuchowski N., Babb J. S., Salganicoff M., Naidich D. P., Anand V., Park S., Vlahos I., and Ko J. P. Benefit of computer-aided detection analysis for the detection of subsolid and solid lung nodules on thin- and thick-section CT. *AJR Am J Roentgenol*, 200:74–83, 2013.

[25] Brown M. S., Goldin J. G., Rogers S., Kim H. J., Suh R. D., McNitt-Gray M. F., Shah S. K., Truong D., Brown K., Sayre J. W., Gjertson D. W., Batra P., and Aberle D. R. Computer-aided lung nodule detection in CT: Results of large-scale observer test. *Acad Radiol*, 12:681–686, 2005.

[26] Brown M. S., Goldin J. G., Suh R. D., McNitt-Gray M. F., Sayre J. W., and Aberle D. R. Lung micronodules: automated method for detection at thin-section CT — initial experience. *Radiology*, 226:256–262, 2003.

[27] Wormanns D., Fiebich M., Saida M., Diederich S., and Heindel W. Automatic detection of pulmonary nodules at spiral CT: clinical application of a computer-aided diagnosis system. *Eur Radiol*, 12:1052–1057, 2002.

[28] MacMahon H. Improvement in detection of pulmonary nodules: Digital image processing and computer-aided diagnosis. *Radiographics*, 20:1169–1177, 2000.

[29] de Koning H. J., Meza R., Plevritis S. K., Ten Haaf K., Munshi V. N., Jeon J., Erdogan S. A., Kong C. Y., Han S. S., van Rosmalen J., Choi S. E., Pinsky P. F., de Gonzalez A. B., Berg C. D., Black W. C., Tammemägi M. C., Hazelton W. D., Feuer E. J., and McMahon P. M. Benefits and harms of computed tomography lung cancer screening strategies: A comparative modeling study for the U.S. preventive services task force. *Ann Intern Med*, 365:311–320, 2013.

[30] Black W. C., Gareen I. F., Soneji S. S., Sicks J. D., Keeler E. B., Aberle D. R., Naeim A., Church T. R., Silvestri G. A., Gorelick J., Gatsonis C., and N. L. S. T. R. T. Cost-effectiveness of CT screening in the National Lung Screening Trial. *N Engl J Med*, 371:1793–1802, 2014.

[31] Aberle D. R., Henschke C. I., McLoud T. C., and Boiselle P. M. Expert opinion: Barriers to CT screening for lung cancer. *J Thorac Imaging*, 27:208, 2012.

[32] Prokop M. Lung cancer screening: The radiologist's perspective. *Semin Respir Crit Care Med*, 35:91–98, 2014.

[33] Cohen J. A coefficient of agreement for nominal scales. *Educ Psychol Meas*, 20:37–46, 1960.

[34] Tan M., Deklerck R., Jansen B., Bister M., and Cornelis J. A novel computer-aided lung nodule detection system for CT images. *Med Phys*, 38:5630–5645, 2011.

[35] Messay T., Hardie R. C., and Rogers S. K. A new computationally efficient CAD system for pulmonary nodule detection in CT imagery. *Med Image Anal*, 14:390–406, 2010.

[36] Murphy K., van Ginneken B., Schilham A. M. R., de Hoop B. J., Gietema H. A., and Prokop M. A large scale evaluation of automatic pulmonary nodule detection in chest CT using local image features and k-nearest-neighbour classification. *Med Image Anal*, 13:757–770, 2009.

[37] Sousa J. R., Silva A. C., de Paiva A. C., and Nunes R. A. Methodology for automatic detection of lung nodules in computerized tomography images. *Comput Methods Programs Biomed*, 98: 1–14, 2009.

[38] Ye X., Lin X., Dehmeshki J., Slabaugh G., and Beddoe G. Shape-based computer-aided detection of lung nodules in thoracic CT images. *IEEE Trans Biomed Eng*, 56:1810–1820, 2009.

[39] Li Q., Li F., and Doi K. Computerized detection of lung nodules in thin-section CT images by use of selective enhancement filters and an automated rule-based classifier. *Acad Radiol*, 15: 165–175, 2008.

[40] Retico A., Delogu P., Fantacci M. E., Gori I., and Martinez A. P. Lung nodule detection in low-dose and thin-slice computed tomography. *Comput Biol Med*, 38:525–534, 2008.

[41] Bellotti R., Carlo F. D., Gargano G., Tangaro S., Cascio D., Catanzariti E., Cerello P., Cheran S. C., Delogu P., Mitri I. D., Fulcheri C., Grosso D., Retico A., Squarcia S., Tommasi E., and Golosio B. A CAD system for nodule detection in low-dose lung CTs based on region growing and a new active contour model. *Med Phys*, 34:4901–4910, 2007.

[42] Dehmeshki J., Ye X., Lin X., Valdivieso M., and Amin H. Automated detection of lung nodules in CT images using shape-based genetic algorithm. *Comput Med Imaging Graph*, 31:408–417, 2007.

[43] Enquobahrie A. A., Reeves A. P., Yankelevitz D. F., and Henschke C. I. Automated detection of

small pulmonary nodules in whole lung CT scans. *Acad Radiol*, 14:579–593, 2007.

[44] Marten K. and Engelke C. Computer-aided detection and automated CT volumetry of pulmonary nodules. *Eur Radiol*, 17:888–901, 2007.

[45] Jacobs C., Sánchez C. I., Saur S. C., Twellmann T., de Jong P. A., and van Ginneken B. Computer-aided detection of ground glass nodules in thoracic CT images using shape, intensity and context features. In *Med Image Comput Comput Assist Interv*, volume 6893 of *Lect Notes Comput Sci*, pages 207–214, 2011.

[46] Tao Y., Lu L., Dewan M., Chen A. Y., Corso J., Xuan J., Salganicoff M., and Krishnan A. Multi-level ground glass nodule detection and segmentation in CT lung images. In *Med Image Comput Comput Assist Interv*, volume 5762 of *Lect Notes Comput Sci*, pages 715–723, 2009.

[47] Ye X., Lin X., Beddoe G., and Dehmeshki J. Efficient computer-aided detection of ground-glass opacity nodules in thoracic CT images. In *Conf Proc IEEE Eng Med Biol Soc*, pages 4449–4452, 2007.

[48] Zhou J., Chang S., Metaxas D. N., Zhao B., Ginsberg M. S., and Schwartz L. H. An automatic method for ground glass opacity nodule detection and segmentation from CT studies. In *Conf Proc IEEE Eng Med Biol Soc*, pages 3062–3065, 2006.

[49] Kim K. G., Goo J. M., Kim J. H., Lee H. J., Min B. G., Bae K. T., and Im J.-G. Computer-aided diagnosis of localized ground-glass opacity in the lung at CT: Initial experience. *Radiology*, 237: 657–661, 2005.

[50] Naidich D. P., Bankier A. A., MacMahon H., Schaefer-Prokop C. M., Pistolesi M., Goo J. M., Macchiarini P., Crapo J. D., Herold C. J., Austin J. H., and Travis W. D. Recommendations for the management of subsolid pulmonary nodules detected at CT: a statement from the Fleischner Society. *Radiology*, 266:304–317, 2013.

[51] Godoy M. C. B. and Naidich D. P. Subsolid pulmonary nodules and the spectrum of peripheral adenocarcinomas of the lung: recommended interim guidelines for assessment and management. *Radiology*, 253:606–622, 2009.

[52] van Rikxoort E. M., de Hoop B., Viergever M. A., Prokop M., and van Ginneken B. Automatic lung segmentation from thoracic computed tomography scans using a hybrid approach with error detection. *Med Phys*, 36:2934–2947, 2009.

[53] van Ginneken B., Baggerman W., and van Rikxoort E. M. Robust segmentation and anatomical labeling of the airway tree from thoracic CT scans. In *Med Image Comput Comput Assist Interv*, volume 5241 of *Lect Notes Comput Sci*, pages 219–226, 2008.

[54] van Dongen E. and van Ginneken B. Automatic segmentation of pulmonary vasculature in thoracic CT scans with local thresholding and airway wall removal. In *IEEE Int Symp Biomedical Imaging*, pages 668–671, 2010.

[55] Heitmann K. R., Kauczor H., Mildenberger P., Uthmann T., Perl J., and Thelen M. Automatic detection of ground glass opacities on lung HRCT using multiple neural networks. *Eur Radiol*, 7:1463–1472, 1997.

[56] Kauczor H.-U., Heitmann K., Heussel C. P., Marwede D., Uthmann T., and Thelen M. Automatic detection and quantification of ground-glass opacities on high-resolution CT using multiple neural networks: Comparison with a density mask. *AJR Am J Roentgenol*, 175:1329–1334,

2000.

[57] Kuhnigk J. M., Dicken V., Bornemann L., Bakai A., Wormanns D., Krass S., and Peitgen H. O. Morphological segmentation and partial volume analysis for volumetry of solid pulmonary lesions in thoracic CT scans. *IEEE Trans Med Imaging*, 25:417–434, 2006.

[58] Hu M.-K. Visual pattern recognition by moment invariants, computer methods in image analysis. *IRE Trans Inf Theory*, 8:179–187, 1962.

[59] Frangi A. F., Niessen W. J., Vincken K. L., and Viergever M. A. Multiscale vessel enhancement filtering. In *Med Image Comput Comput Assist Interv*, volume 1496 of *Lect Notes Comput Sci*, pages 130–137, 1998.

[60] Ojala T., Pietikainen M., and Maenpaa T. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Trans Pattern Anal Mach Intell*, 24:971–987, 2002.

[61] Sörensen L., Shaker S. B., and de Bruijne M. Texture classification in lung CT using local binary patterns. In *Med Image Comput Comput Assist Interv*, volume 11 of *Lect Notes Comput Sci*, pages 934–941, 2008.

[62] Fukunaga K. *Introduction to Statistical Pattern Recognition*. Academic Press, 2nd edition, 1990.

[63] Pudil P., Novovicova J., and Kittler J. Floating search methods in feature selection. *Pattern Recognit Lett*, 15:1119–1125, 1994.

[64] Jobson J. D. *Applied Multivariate Data Analysis*. Springer-Verlag, 1992.

[65] Cover T. and Hart P. Nearest neighbor pattern classification. *IEEE Trans Inf Theory*, 13:21–27, 1967.

[66] Breiman L. Random forests. *Machine Learning*, 45:5–32, 2001.

[67] Friedman J., Hastie T., and Tibshirani R. Additive logistic regression: A statistical view of boosting. *Ann Stat*, 28:337–374, 2000.

[68] Vapnik V. *The nature of statistical learning theory*. Springer-Verlag, New York, 1995.

[69] van Ginneken B., Armato S. G., de Hoop B., van de Vorst S., Duindam T., Niemeijer M., Murphy K., Schilham A. M. R., Retico A., Fantacci M. E., Camarlinghi N., Bagagli F., Gori I., Hara T., Fujita H., Gargano G., Belloti R., Carlo F. D., Megna R., Tangaro S., Bolanos L., Cerello P., Cheran S. C., Torres E. L., and Prokop M. Comparing and combining algorithms for computer-aided detection of pulmonary nodules in computed tomography scans: the ANODE09 study. *Med Image Anal*, 14:707–722, 2010.

[70] Armato S. G., McLennan G., Bidaut L., McNitt-Gray M. F., Meyer C. R., Reeves A. P., Zhao B., Aberle D. R., Henschke C. I., Hoffman E. A., Kazerooni E. A., MacMahon H., Beek E. J. R. V., Yankelevitz D., Biancardi A. M., Bland P. H., Brown M. S., Engelmann R. M., Laderach G. E., Max D., Pais R. C., Qing D. P. Y., Roberts R. Y., Smith A. R., Starkey A., Batrah P., Caligiuri P., Farooqi A., Gladish G. W., Jude C. M., Munden R. F., Petkovska I., Quint L. E., Schwartz L. H., Sundaram B., Dodd L. E., Fenimore C., Gur D., Petrick N., Freymann J., Kirby J., Hughes B., Casteele A. V., Gupte S., Sallamm M., Heath M. D., Kuhn M. H., Dharaiya E., Burns R., Fryd D. S., Salganicoff M., Anand V., Shreter U., Vastagh S., and Croft B. Y. The Lung Image Database Consortium (LIDC) and Image Database Resource Initiative (IDRI): a completed

reference database of lung nodules on CT scans. *Med Phys*, 38:915–931, 2011.

[71] Peldschus K., Herzog P., Wood S. A., Cheema J. I., Costello P., and Schoepf U. J. Computer-aided diagnosis as a second reader: spectrum of findings in CT studies of the chest interpreted as normal. *Chest*, 128:1517–1523, 2005.

[72] Sánchez C. I., Niemeijer M., Isgum I., Dumitrescu A. V., Suttorp-Schulten M. S. A., Abràmoff M. D., and van Ginneken B. Contextual computer-aided detection: Improving bright lesion detection in retinal images and coronary calcification identification in CT scans. *Med Image Anal*, 16:50–62, 2012.

[73] Song Y., Cai W., Kim J., and Feng D. A multi-stage discriminative model for tumor and lymph node detection in thoracic images. *IEEE Trans Med Imaging*, 31:1061–1075, 2012.

[74] Hupse R. and Karssemeijer N. The use of contextual information for computer aided detection of masses in mammograms. In *Medical Imaging*, volume 7260 of *Proceedings of the SPIE*, page 72600Q, 2009.

[75] Rees D. and Murray J. Silica, silicosis and tuberculosis. *Int J Tuberc Lung Dis*, 11:474–484, 2007.

[76] Mets O. M., Rooyackers J., van Amelsvoort-van de Vorst S., Mali W. P., de Jong P. A., and Prokop M. Increased micronodule counts are more common in occupationally silica dust-exposed smokers than in control smokers. *J Occup Environ Med*, 2012.

[77] Jacobs C., van Rikxoort E. M., Twellmann T., Scholten E. T., de Jong P. A., Kuhnigk J. M., Oudkerk M., de Koning H. J., Prokop M., Schaefer-Prokop C., and van Ginneken B. Automatic detection of subsolid pulmonary nodules in thoracic computed tomography images. *Med Image Anal*, 18:374–384, 2014.

[78] Lee Y., Hara T., Fujita H., Itoh S., and Ishigaki T. Automated detection of pulmonary nodules in helical CT images based on an improved template-matching technique. *IEEE Trans Med Imaging*, 20:595–604, 2001.

[79] Armato S. G., McLennan G., McNitt-Gray M. F., Meyer C. R., Yankelevitz D., Aberle D. R., Henschke C. I., Hoffman E. A., Kazerooni E. A., MacMahon H., Reeves A. P., Croft B. Y., and Clarke L. P. Lung Image Database Consortium: Developing a resource for the medical imaging research community. *Radiology*, 232:739–748, 2004.

[80] Chakraborty D. P. and Berbaum K. S. Observer studies involving detection and localization: modeling, analysis, and validation. *Med Phys*, 31:2313–2330, 2004.

[81] Efron B. Bootstrap methods: Another look at the jackknife. *Ann Stat*, 7:1–26, 1979.

[82] Armato S. G., McNitt-Gray M. F., Reeves A. P., Meyer C. R., McLennan G., Aberle D. R., Kazerooni E. A., MacMahon H., van Beek E. J. R., Yankelevitz D., Hoffman E. A., Henschke C. I., Roberts R. Y., Brown M. S., Engelmann R. M., Pais R. C., Piker C. W., Qing D., Kocherginsky M., Croft B. Y., and Clarke L. P. The Lung Image Database Consortium (LIDC): an evaluation of radiologist variability in the identification of lung nodules on CT scans. *Acad Radiol*, 14: 1409–1421, 2007.

[83] Drew T., Vo M. L.-H., Olwal A., Jacobson F., Seltzer S. E., and Wolfe J. M. Scanners and drillers: characterizing expert visual search through volumetric images. *J Vis*, 13, 2013.

[84] Camarlinghi N. Automatic detection of lung nodules in computed tomography images: train-

ing and validation of algorithms using public research databases. *EPJ Plus*, 128:1–21, 2013.

[85] Tan M., Deklerck R., Cornelis J., and Jansen B. Phased searching with neat in a time-scaled framework: Experiments on a computer-aided detection system for lung nodules. *Artif Intell Med*, 59:157–67, 2013.

[86] American Cancer Society. Cancer facts and figures 2012, 2012.

[87] MacMahon H., Austin J. H. M., Gamsu G., Herold C. J., Jett J. R., Naidich D. P., Patz E. F., Swensen S. J., and the Fleischner Society. Guidelines for management of small pulmonary nodules detected on CT scans: a statement from the Fleischner Society. *Radiology*, 237:395–400, 2005.

[88] McWilliams A., Tammemagi M. C., Mayo J. R., Roberts H., Liu G., Soghrati K., Yasufuku K., Martel S., Laberge F., Gingras M., Atkar-Khattra S., Berg C. D., Evans K., Finley R., Yee J., English J., Nasute P., Goffin J., Puksa S., Stewart L., Tsai S., Johnston M. R., Manos D., Nicholas G., Goss G. D., Seely J. M., Amjadi K., Tremblay A., Burrowes P., MacEachern P., Bhatia R., Tsao M.-S., and Lam S. Probability of cancer in pulmonary nodules detected on first screening CT. *N Engl J Med*, 369:910–919, 2013.

[89] Yildirim A., Ridge C., Boiselle P., Franquet T., Gevenois P., Schaefer-Prokop C., Tack D., and Bankier A. Differentiating subsolid and solid pulmonary nodules on CT: inter- and intraobserver agreement among experienced thoracic radiologists. In *European Congress of Radiology*, 2013.

[90] van Riel S. J., Schaefer-Prokop C. M., van Rikxoort E. M., van Ginneken B., Prokop M., Schalekamp S., Jacobs C., de Jong P. A., Gietema H. A., and Scholten E. T. Impact of section thickness on classification of pulmonary nodules into solid, part-solid, and non-solid: an observer study. In *Annual Meeting of the Radiological Society of North America*, 2013.

[91] Scholten E. T., Mali W. P. T. M., Prokop M., van Ginneken B., Glandorf R., van Klaveren R., Oudkerk M., and de Jong P. A. Non-solid lung nodules on low-dose computed tomography: comparison of detection rate between 3 visualization techniques. *Cancer Imaging*, 13:150–154, 2013.

[92] Travis W. D., Brambilla E., Noguchi M., Nicholson A. G., Geisinger K. R., Yatabe Y., Beer D. G., Powell C. A., Riely G. J., Van Schil P. E., Garg K., Austin J. H. M., Asamura H., Rusch V. W., Hirsch F. R., Scagliotti G., Mitsudomi T., Huber R. M., Ishikawa Y., Jett J., Sanchez-Cespedes M., Sculier J.-P., Takahashi T., Tsuboi M., Vansteenkiste J., Wistuba I., Yang P.-C., Aberle D., Brambilla C., Flieder D., Franklin W., Gazdar A., Gould M., Hasleton P., Henderson D., Johnson B., Johnson D., Kerr K., Kuriyama K., Lee J. S., Miller V. A., Petersen I., Roggli V., Rosell R., Saijo N., Thunnissen E., Tsao M., and Yankelewitz D. International association for the study of lung cancer/american thoracic society/european respiratory society international multidisciplinary classification of lung adenocarcinoma. *J Thorac Oncol*, 6:244–285, 2011.

[93] Takashima S., Maruyama Y., Hasegawa M., Yamanda T., Honda T., Kadoya M., and Sone S. Prognostic significance of high-resolution CT findings in small peripheral adenocarcinoma of the lung: a retrospective study on 64 patients. *Lung Cancer*, 36:289–295, 2002.

[94] Miao X.-H., Yao Y.-W., Yuan D.-M., Lv Y.-L., Zhan P., Lv T.-F., Liu H.-B., and Song Y. Prognostic value of the ratio of ground glass opacity on computed tomography in small lung adenocarcinoma: A meta-analysis. *J Thorac Dis*, 4:265–271, 2012.

[95] Raad R. A., Suh J., Harari S., Naidich D. P., Shiau M., and Ko J. P. Nodule characterization: subsolid nodules. *Radiol Clin North Am*, 52:47–67, 2014.

[96] Jeon K. N., Goo J. M., Lee C. H., Lee Y., Choo J. Y., Lee N. K., Shim M.-S., Lee I. S., Kim K. G., Gierada D. S., and Bae K. T. Computer-aided nodule detection and volumetry to reduce variability between radiologists in the interpretation of lung nodules at low-dose screening computed tomography. *Invest Radiol*, 47:457–461, 2012.

[97] Lee K. H., Goo J. M., Park S. J., Wi J. Y., Chung D. H., Go H., Park H. S., Park C. M., and Lee S. M. Correlation between the size of the solid component on thin-section CT and the invasive component on pathology in small lung adenocarcinomas manifesting as ground-glass nodules. *J Thorac Oncol*, 9:74–82, 2014.

[98] Maldonado F., Boland J. M., Raghunath S., Aubry M. C., Bartholmai B. J., Deandrade M., Hartman T. E., Karwoski R. A., Rajagopalan S., Sykes A.-M., Yang P., Yi E. S., Robb R. A., and Peikert T. Noninvasive characterization of the histopathologic features of pulmonary nodules of the lung adenocarcinoma spectrum using computer-aided nodule assessment and risk yield (CANARY)–a pilot study. *J Thorac Oncol*, 8:452–460, 2013.

[99] Kawata Y., Niki N., Ohmatsu H., Kusumoto M., Tsuchida T., Eguchi K., Kaneko M., and Moriyama N. Quantitative classification based on CT histogram analysis of non-small cell lung cancer: correlation with histopathological characteristics and recurrence-free survival. *Med Phys*, 39:988–1000, 2012.

[100] Yanagawa M., Tanaka Y., Leung A. N., Morii E., Kusumoto M., Watanabe S., Watanabe H., Inoue M., Okumura M., Gyobu T., Ueda K., Honda O., Sumikawa H., Johkoh T., and Tomiyama N. Prognostic importance of volumetric measurements in stage I lung adenocarcinoma. *Radiology*, 272:557–567, 2014.

[101] Rühaak J., Heldmann S., Kipshagen T., and Fischer B. Highly accurate fast lung CT registration. In *Medical Imaging*, volume 8669 of *Proceedings of the SPIE*, page 86690Y, 2013.

[102] Murphy K., van Ginneken B., Reinhardt J. M., Kabus S., Ding K., Deng X., Cao K., Du K., Christensen G. E., Garcia V., Vercauteren T., Ayache N., Commowick O., Malandain G., Glocker B., Paragios N., Navab N., Gorbunova V., Sporring J., de Bruijne M., Han X., Heinrich M. P., Schnabel J. A., Jenkinson M., Lorenz C., Modat M., McClelland J. R., Ourselin S., Muenzing S. E. A., Viergever M. A., Nigris D. D., Collins D. L., Arbel T., Peroni M., Li R., Sharp G. C., Schmidt-Richberg A., Ehrhardt J., Werner R., Smeets D., Loeckx D., Song G., Tustison N., Avants B., Gee J. C., Staring M., Klein S., Stoel B. C., Urschler M., Werlberger M., Vandemeulebroucke J., Rit S., Sarrut D., and Pluim J. P. W. Evaluation of registration methods on thoracic CT: The EMPIRE10 challenge. *IEEE Trans Med Imaging*, 31:1901–1920, 2011.

[103] Aoki T., Murakami S., Kim H., Fujii M., Takahashi H., Oki H., Hayashida Y., Katsuragawa S., Shiraishi J., and Korogi Y. Temporal subtraction method for lung nodule detection on successive thoracic CT soft-copy images. *Radiology*, 271:255–261, 2014.

[104] Staring M., Pluim J. P. W., de Hoop B. J., Klein S., van Ginneken B., Gietema H., Nossent G., Schaefer-Prokop C. M., van de Vorst S., and Prokop M. Image subtraction facilitates assessment of volume and density change in ground-glass opacities in chest CT. *Invest Radiol*, 44:61–66, 2009.

[105] Maastricht University, the Netherlands. Article 23 of the Regulation governing the attainment

of doctoral degrees, 2013.

[106] Kuhnigk J.-M., Dicken V., Zidowitz S., Bornemann L., Kuemmerlen B., Krass S., Peitgen H.-O., Yuval S., Jend H.-H., Rau W. S., and Achenbach T. New tools for computer assistance in thoracic CT part 1. Functional analysis of lungs, lung lobes and bronchopulmonary segments. *Radiographics*, 25:525–536, 2005.

[107] van Rikxoort E. M., de Hoop B., van de Vorst S., Prokop M., and van Ginneken B. Automatic segmentation of pulmonary segments from volumetric chest CT scans. *IEEE Trans Med Imaging*, 28:621–630, 2009.

[108] van Rikxoort E. M., Prokop M., de Hoop B., Viergever M. A., Pluim J. P. W., and van Ginneken B. Automatic segmentation of the pulmonary lobes from fissures, airways, and lung borders: evaluation of robustness against missing data. In *Med Image Comput Comput Assist Interv*, volume 12 of *Lect Notes Comput Sci*, pages 263–271, 2009.

[109] McWilliams A., Mayo J., MacDonald S., leRiche J. C., Palcic B., Szabo E., and Lam S. Lung cancer screening: A different paradigm. *Am J Respir Crit Care Med*, 168:1167–1173, 2003.

[110] de Hoop B., Gietema H., van de Vorst S., Murphy K., van Klaveren R. J., and Prokop M. Pulmonary ground-glass nodules: increase in mass as an early indicator of growth. *Radiology*, 255: 199–206, 2010.

[111] American College of Radiology. Lung CT screening reporting and data system (Lung-RADS), 2014.

[112] van Ginneken B., Jacobs C., Scholten E. T., Prokop M., and de Jong P. A. Feasibility of rapid reading of CT lung cancer screening with computer-aided detection support. In *Annual Meeting of the Radiological Society of North America*, 2014.

[113] Isgum I., Prokop M., Niemeijer M., Viergever M., and van Ginneken B. Automatic coronary calcium scoring in low-dose chest computed tomography. *IEEE Trans Med Imaging*, 31:2322 – 2334, 2012.

[114] van Ginneken B., Setio A. A. A., Jacobs C., and Ciompi F. Off-the-shelf convolutional neural network features for pulmonary nodule detection in computed tomography scans. In *IEEE Int Symp Biomedical Imaging*, pages 286–289, 2015.

[115] Scholten E. T., Horeweg N., de Koning H. J., Vliegenthart R., Oudkerk M., Mali W. P. T. M., and de Jong P. A. Computed tomographic characteristics of interval and post screen carcinomas in lung cancer screening. *Eur Radiol*, 25:81–88, 2015.

[116] Gilbert F. J., Astley S. M., Gillan M. G. C., Agbaje O. F., Wallis M. G., James J., Boggis C. R. M., Duffy S. W., and CADET II Group. Single reading with computer-aided detection for screening mammography. *N Engl J Med*, 359:1675–1684, 2008.

[117] Fenton J. J., Abraham L., Taplin S. H., Geller B. M., Carney P. A., D'Orsi C., Elmore J. G., and Barlow W. E. Effectiveness of computer-aided detection in community mammography practice. *J Natl Cancer Inst*, 103:1152–1161, 2011.

Dankwoord

Als laatste wil ik iedereen bedanken die heeft bijgedragen aan de totstandkoming van dit proefschrift.

Prof. dr. van Ginneken, beste Bram, ik wil je ontzettend bedanken voor je begeleiding tijdens mijn promotieonderzoek. Ik heb enorm veel van je geleerd. De betrokkenheid en passie die jij elke dag weer toont voor het onderzoek binnen DIAG is uitzonderlijk. Ik vind het erg knap dat je zoveel projecten tegelijk weet te runnen en daarbij nooit een detail lijkt te vergeten. Ook is het erg prettig een begeleider te hebben die het grote doel altijd goed voor ogen heeft en bijstuurt waar nodig. Door het verkrijgen van jouw VICI subsidie kreeg ons onderzoek omtrent automatisering van longkanker screening een enorme boost waar ik nu nog steeds de vruchten van pluk.

Prof. dr. Schaefer-Prokop, beste Cornelia, jouw klinische input is erg belangrijk geweest tijdens mijn promotietraject. Je hebt een oprechte passie voor onderzoek die me erg kan inspireren. Ik wil je ook bedanken voor jouw warme persoonlijkheid en oprechte gastvrijheid. Het is prettig om met jou samen te werken.

Dr. van Rikxoort, beste Eva, ik ben erg blij dat jij mijn directe begeleider bent geweest. Ik denk dat we het erg goed kunnen vinden en weinig woorden nodig hebben om te bepalen hoe we problemen binnen het chest CT onderzoek aanpakken. Je bent een vrolijk en positief ingesteld persoon bij wie ik altijd kon aankloppen met vragen. Ook bedankt voor alle gezelligheid tijdens conferenties, borrels, etc. Als laatste wil ik je alle geluk en succes wensen met je eigen bedrijf.

Prof. dr. Prokop, beste Mathias, je staat niet als promotor of copromotor op mijn proefschrift maar toch heb je erg veel invloed gehad op mijn onderzoek. Door de wekelijkse pulmo meetings heb ik vaak met jou van gedachten kunnen wisselen. Jouw inzichten en klinische kennis zijn van onschatbare waarde voor elk onderzoek binnen radiologie. Dankzij jouw natuurkunde achtergrond ben je ook in staat om methodologisch commentaar op de ontwikkelde systemen te geven. We hebben geluk dat we jou als hoofd van onze afdeling hebben.

Prof. dr. Brown, dear Matt, I would like to thank you for the opportunity to do an internship at UCLA in the summer of 2013, and for participating in the manuscript committee of this thesis. I enjoyed my stay in your group a lot and you really made me feel at home during my visit. I look forward to discuss the contents of my thesis with you during my defense. Prof. Dr. Aberle, dear Denise, thank you very much for your help with my project at UCLA. I enjoyed working with you and thank you for our useful discussions. I hope we can collaborate again in the near future. Furthermore, I would like to thank dr. Goldin, dr. McNitt-Gray, William, Sutida, Andrea, Pechin, Greg, Stefano, Heidi, Danny, Andy, Katherine, Hak, Karen, Erin, Eloisa, and everyone else in the lab for my time at UCLA.

Dr. Platel, beste Bram, ik wil jou ook erg bedanken want dankzij jou ben ik bij DIAG terecht gekomen. Wij zijn samen als eerste aangesteld bij Fraunhofer MEVIS en hebben samen ook een aantal tripjes naar Bremen gemaakt. Ik ken je al sinds mijn tijd op de TU Eindhoven en ik denk dat wij het altijd erg goed hebben kunnen vinden. Aangezien ik in dezelfde metropool ben geboren waar jij bent opgegroeid, is dat misschien ook niet zo gek.

Beste Sjoerd Opdam, bedankt voor het onderzoek wat je gedaan hebt als student binnen onze groep. Jouw project is de basis voor hoofdstuk 4 van dit proefschrift geweest.

Dr. de Jong, beste Pim, bedankt voor jouw bijdrage aan de artikelen in dit proefschrift.

I would like to thank my supervisors from MeVis Medical Solutions which helped me during the years. Thorsten Twellmann, Stephan Fromme, and Andreas Beck, thank you very much for your contributions to my work. A special thanks to Stefan Saur who really helped with the design and implementation of the system described in chapter 2 of this thesis.

To my colleagues at Fraunhofer MEVIS, in particular Horst Hahn, Jan Rühaak, Jan-Martin Kuhnigk, Bianca Lassen, and Michael Schmidt: thank you all for your help. Fraunhofer MEVIS is a great research institute and I very much enjoyed my visits to Bremen. Also, it was an great honor to participate in the prestigious Fraunhofer soccer tournament.

When I started my PhD project, I was the only person working on lung nodules within DIAG. Now, we have a fairly large team around this topic. Francesco, Arnaud, Sarah, Kaman, Paul, it is really a pleasure to work with you. Thanks a lot for that. I also would like to mention the other two chest CT researchers in our group with whom I had a lot of meetings: Leticia and Jean-Paul. Thanks for listening to all the nodule talks during our chest CT meetings. Een speciaal dankwoord ook voor Ernst Scholten die me ontzettend heeft geholpen. Ik prijs me gelukkig dat ik veel met een ervaren radioloog heb kunnen samenwerken. Verder vind ik het fantastisch dat je nu een aanstelling in onze groep hebt en onderdeel van ons team bent.

Beste Christian Mol, bedankt voor jouw hulp tijdens het eerste jaar. Je hebt me erg geholpen bij het leren van C++ en het begrijpen van de codebase in DIAG. Laten we snel weer eens een potje gaan squashen.

Ik wil graag alle onderzoekers in DIAG bedanken. DIAG is een erg prettige onderzoeksgroep waar je als beginnend onderzoeker veel kan leren. Naast werk zijn er ook allerlei sociale activiteiten zoals borrels, feestjes en uiteraard de onvergetelijke DIAG weekenden. Een speciaal dankwoord voor ons ochtendkoffie team waarbij de harde kern nog steeds met de kenmerkende Paddington mokken rondloopt. Ons

sociale kwartiertje in de ochtend zorgt altijd voor een goede start van de dag. De volgende namen mogen zeker niet in dit dankwoord ontbreken omdat ze op hun eigen manier invloed hebben gehad op mij en/of mijn onderzoek: Albert, Clarisa, Geert, Jan, Laurens, Mark, Pragnya, Rick, Rieneke, Steven, Suzan en Wendy, bedankt!

Beste Jan-Jurre, ik vind het ongelofelijk dat je na vier jaar mijn privéchauffeur te zijn geweest ook nog akkoord bent gegaan mijn paranimf te zijn. Waar heb ik dat aan te danken? Ik zit natuurlijk maar wat te geinen. Kort nadat je bij DIAG begon, vroeg je me of het geen goed idee was om samen te gaan carpoolen. Alhoewel carpoolen natuurlijk heel leuk klinkt, vereist het wel degelijk flexibiliteit van beiden. Ik vind het geweldig dat wij dit al vier jaar moeiteloos volhouden. We zijn ondertussen vrienden geworden en ik vind het top dat je mijn paranimf bent.

Beste Nick, lieve broer, al sinds we erg klein zijn (waarschijnlijk sinds jouw geboorte) doen we heel veel samen: voetballen, tennissen, wielrennen, zelfde vriendengroep, bootje opknappen en ga zo maar door. Je zou bijna zeggen dat wij elkaar teveel zien en dat dit niet goed kan gaan, maar niets is minder waar want we hebben praktisch nooit onenigheid. Ik ben erg blij en trots met jou als mijn broer. Bedankt dat je mijn paranimf wilt zijn.

Ik wil graag ook mijn vriendengroep bedanken. Cas, Koen, Menno, Nard, Sjoerd en Yori, het is top om een erg hechte vriendengroep te hebben. We kennen elkaar al erg lang en alhoewel we momenteel best verspreid wonen, zien we elkaar nog zeer regelmatig. Jullie kritische vragen (Colin, heb je dat longkanker probleem nu al opgelost? Colin, heb je dat artikel nu nog niet gepubliceerd?) brachten telkens weer die nieuwe invalshoek op mijn onderzoek die ik nodig had. Ook een speciaal woord van dank voor Bart, die tijdens mijn eerste jaar elke ochtend voor verse koffie en gezelligheid in de auto zorgde.

Lindsey, lieve zus, ik wil ook even zeggen dat ik erg trots ben op jou. Ondanks dat het momenteel wat onstuimig is in de modebranche, vind jij vast snel een leuke vaste baan. Je bent een topper!

Ik wil ook graag de rest van mijn familie bedanken. Elsemieke, bedankt dat je zo'n fijne schoonzus bent. Leo, Rob, Lars, Yvonne, Desiree en Dien de Beer, bedankt dat jullie me altijd zo thuis laten voelen in jullie familie.

Lieve pap en mam, heel erg bedankt voor alles. Uiteraard had ik dit zonder jullie nooit kunnen doen. Jullie onvoorwaardelijke steun voelt voor mij bijna vanzelfsprekend, maar ik realiseer me dat het dat zeker niet is. Ik vind het altijd heel fijn om thuis te komen.

Als laatste natuurlijk mijn lieve Kim. Heel erg bedankt voor jouw steun de afgelopen jaren. Ik ben ontzettend gelukkig met jou en kan me geen leven zonder jou voorstellen. Je bent de beste!

Curriculum Vitae

Colin Jacobs was born in Boxtel, the Netherlands, on August 24th, 1986. In 2004, he started the Biomedical Engineering program at the Eindhoven University of Technology (TU/e), in Eindhoven, the Netherlands. His master's thesis project entitled "A feasibility study to reconstruct the optic radiation for epilepsy surgery using fMRI-seeded DTI tractography" was a collaboration project between the TU/e and epileptology and sleep medicine expertise centre Kempenhaeghe. In October 2010, he started as a PhD student in the Diagnostic Image Analysis Group in the Department of Radiology, Radboud University Medical Center in Nijmegen, the Netherlands. His project is conducted in close collaboration with research institute Fraunhofer MEVIS, where Colin also had a position as Biomedical Engineer between October 2010 and February 2013. His thesis project was funded by MeVis Medical Solutions AG, Bremen, Germany. The results of the work he carried out in the Diagnostic Image Analysis Group are described in this thesis.